Relationship Material: Using Machine Learning to Identify Variables of Importance that Best Predict Lifestyle Choice in the Add Health Longitudinal Dataset

Matthew J. Domiteaux

Charlottesville, Virginia

Bachelor of Arts, University of Texas, 2012

A Dissertation Presented to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Doctor of Philosophy

Department of Psychology

University of Virginia

July, 2020

Abstract

Machine learning has been increasing in popularity due to its potential to provide major insights into a variety of complex topics. However, the applications of these techniques to the study of psychology is not yet widespread. This study seeks to use a specific type of supervised machine learning - multi-class classification - to predict who marries, cohabitates, or remains single by young adulthood (i.e., ages 24 to 32). This study applied machine learning to an extensive dataset, the National Longitudinal Study of Adolescent to Adult Health (Add Health), which is a rich, longitudinal survey that includes a diverse sample of over ten thousand participants and several thousands of variables collected in five waves spanning two decades. Variables within Add Health tap dozens of psychological and behavioral constructs that may serve as predictors of lifestyle choice. Broadly stated this study examined: 1) How well can marriage, cohabitation, and singlehood be predicted within the Add Health dataset?; 2) Do certain topic constructs in Add Health such as substance use or personality influence these predictions more so than others among the widest range of predictors possible?; 3) Are there variables from earlier in the lifespan that can accurately predict outcomes that occur later in life up to young adulthood? In order to answer these questions this study applied and compared the results from multiple machine learning models using a sophisticated, multi-model, cross-validation approach. The major implications of this study are twofold: 1) uncover which variables are most important and predictive when it comes to lifestyle choice in young adulthood; and 2) provide a template for

using machine learning in the context of large datasets that can be applied to other research questions and outcomes (e.g., body mass index [BMI], intelligence).

Keywords: relationships, marriage, cohabitation, singlehood, machine learning, predictive analysis, Add Health

Acknowledgements

I would like to acknowledge each member of my dissertation committee for their unique contributions to my journey as a graduate student and as a clinician. First of all, I want to thank Dr. Eric Turkheimer, who taught me how to improve my skills as a scientific thinker and researcher over the years. Secondly, I want to thank Dr. Bob Emery, who taught me about the science behind relationships that later piqued my interest in couples research. Additionally, working with Bob in clinical supervision led to my long-term interest in couples therapy and to my choosing this dissertation topic. Next, I would like to thank Dr. Lee Llewellyn, who has been both a fantastic clinical supervisor and who has offered incredible personal support during the challenges I experienced in graduate school. I would also like to thank Dr. Jeff Boichuk from the Mcintire School of Commerce, who has provided me with advice on how to apply machine learning to my work in a way that made this dissertation possible. Thanks to the help of Dr. Turkheimer, Dr. Emery and Dr. Llewellyn I was able to secure an internship with the Raymond Murphy VA Clinical Psychology Internship program and complete the required clinical training for my Doctoral degree in a challenging environment that enabled me to grow both as a clinician and as a person.

I would additionally like to thank my friends and family who have been there to support me throughout this process, offering help, moral support, and advice. These include but are not limited to David Dobolyi, Rachel Narr, Karl Fua, and McKinley Herndon. I would also like to thank my parents Mark and Lisa Domiteaux, who have been my cheerleaders from the very beginning.

Table of Contents

Relationship Material: Using Machine Learning to Identify Variables of Importance that Best Predict Lifestyle Choice in the Add Health Longitudinal Dataset

Both academic researchers and the media have devoted substantial attention to the fluctuating decline in marriage rates, a significant issue facing the United States (US) population (Nock, 2005). Nock negatively characterized this decline in marriage rates by reviewing the positive associations of marriage and suggested that a reduction in marriage levels could have socially deleterious effects. Much of the existing literature on lifestyle choices have focused on specific outcomes in isolation (e.g., looking at marriage alone or comparing marriage versus cohabitation). In contrast to such studies, the goal of this research is to determine the best model for simultaneously predicting a more complete set of outcomes - including marriage, cohabitation, and singlehood - in an encompassing model that includes the broadest possible range of predictors from a multi-decade, longitudinal dataset. This research aims to provide a clearer picture of how variables from across the lifespan determine lifestyle choice in later life. This work can reveal how early life variables influence an individual's decision to live as a single person, to cohabitate, or to marry and how these variables affect several significant domains relevant to the study of human behavior. The ability to predict an individual's pathway could lead to interventions that could improve their quality and length of life.

Marriage

Past research focused on highlighting the benefits of marriage has motivated policymakers to create interventions encouraging couples to get married and to increase the health and durability of existing marriages (Nock, 2005; Allen et al., 2012). Many policymakers and legislators have taken up the promotion of marriage formation and maintenance as an explicit goal (Lichter, 2001; Schwartz, 2005; Hawkins et al., 2012). These efforts have included

both public and private programs intended to: 1) educate the public on the benefits of marriage (Hawkins, 2013); 2) strengthen existing marriages (Allen et al., 2012); and 3) encourage future marriages (Eryigit et al., 2010; Kerpelman, 2012).

Empirical evaluations of the efficacy of programs aimed at influencing marriage outcomes have shown positive results, yet these programs have not reversed the overall decline in marriage rates in recent times (Stanley, 2001; Hsueh et al., 2012). More specifically, despite the previously described efforts to strengthen existing marriages and increase the number of marriages, the US's marriage rate has consistently fallen across the last several decades (Lundberg et al., 2016; Silva et al., 2016). This decline can be attributed to several factors, all of which are supported across multiple studies. Firstly, the average age at which individuals first get married has increased, suggesting that couples are choosing to delay marriage (Lundberg et al., 2016): According to 2011 US Census data, the median age of first marriage rose from approximately 20 to 23 for women and men, respectively, in the 1950s to 26 to 28 in 2010, respectively (Cohn, 2011). Secondly, there has been an increase in the number of people who never get married (Garrison, 2007). Finally, the increasing focus on gender equality across the US has pushed women to delay marriage in the pursuit of education and careers (Silva et al., 2016; Hill, 2020).

Existing research focused on marriage decline investigated changing patterns in individual's choices, which yielded several significant findings: 1) cohabitation has become increasingly popular (Eickmeyer & Manning, 2018); 2) individuals are choosing education and employment over marriage in some cases while delaying it in others (Isen & Stevenson, 2010; Silva, 2018); and 3) the economic and non-economic factors that influence marriage are shifting (Jamison, 2018; Silva et al., 2016). Perhaps in response to these patterns, some individuals

choose to abandon the pursuit of marriage altogether while focusing on cohabitation (Manning, 2020). Other explorations of the marriage decline phenomena have investigated changes in technology and the economy, including the development of oral contraception, the legalization of abortion, changes in household technology, and the narrowing of the male-female wage gap (Cherlin et al., 2016; Goldin & Katz, 2002; Greenwood & Guner, 2008; Isen & Stevenson, 2010; Silva, 2020). While there are several potential causes for the overall decline in marriage rates, recent research suggests that certain demographics are at a higher risk for never-marrying or getting divorced than others (Schwizer, 2020).

Zimmerman and Easterlin (2006) discussed setpoint theory and its role in marriage and happiness. While researchers using setpoint theory have attempted to understand the role of personality and genetics in happiness and marriage, little research has come due to an inability to form public policy around such findings. As marriage offers many physical and psychological health benefits and economic opportunities, understanding the reasons individuals are choosing to remain unmarried or get divorced should be understood to improve the intervention efforts currently in place (Chin et al., 2017; Zimmerman & Easterlin, 2006).

Cohabitation

Marriage and cohabitation offer many of the same benefits for couples while cohabiting before marriage can improve the benefits of both (Zimmerman & Easterlin, 2006). Other research suggests that marriage alternatives are less beneficial than marriage itself, citing instability as the primary cause (Booth & Johnson, 1988; DeMaris & Rao, 1992; Teachman & Polonko, 1990). However, more recent research on cohabitation contradicts such claims (Amato, 2015; Heikel & Wagner, 2020; Zimmermann & Easterlin, 2006). The most important distinction to be made is the similarities between marriage and cohabitation. Little evidence has been shown

to suggest that cohabitation is less beneficial than marriage, but the majority of research indicates that individuals who cohabit experience a multitude of positive outcomes singletons do not.

Singlehood

Pew Center research from 2014 indicated that a larger percentage of Americans remained unmarried than ever before (Wang & Parker, 2014). While a large and growing number of individuals were expected to be single people, the majority of individuals included in this data would most likely, at some point, cohabitate.

Beyond the negative economic, physiological, and psychological effects of singlehood, a negative cultural stereotyping on singletons exists (DePaulo & Morris, 2006). Life satisfaction amongst singletons is heterogeneous (Tinomen, 2013). This variation, to some extent, can be explained by interpersonal variables like the number of deep, lifelong friendships, similar to familial relations, that a single person maintains. The subjects in Tinomen's study reported varying degrees of life satisfaction that mostly depended upon the formation of friendships maintained throughout life and cited by the participants as highly important to them, suggesting that even singles still rely on interpersonal relationships for life satisfaction. However, some suggest that the quality and quantity of friendships are on the decline (Knox, 2018; Brashears & Brashears, 2015; Bryner, 2011). While the research into the cause of this decline fails to be conclusive in its findings, the long-lasting impact may impact the lives of singletons.

Present Study

Prior research on understanding why individuals have a preference toward marriage, cohabitation, or singlehood has primarily focused on these outcomes in isolation, but an accurate assessment should include individual differences that account for the diverse determinants

involved in lifestyle choice, as suggested in previous studies (e.g., Amato, 2015; Zimmerman & Easterlin, 2006). While studies that attempt to account for these differences exist, their scope has been limited by the use of regression analyses and a focus on a small number of variables. For example, Amato (2015) used a fixed-effects model to explore whether marriages and cohabiting relationships led to observable mental health improvements across multiple waves of the National Longitudinal Study of Adolescent to Adult Health (Add Health; Add Health Codebook Explorer, 2017). This analysis also examined the effect of gender on the outcome, while also including control variables to account for time-independent factors, such as race and time-dependent factors, such as age and education. Studies like Amato's (2015) offer valuable insights into the benefits associated with various lifestyle choices. Nevertheless, the utilization of a relatively small number of Add Health variables makes it difficult to interpret the true underlying effects since only a few variables are included in their models relative to the full scope of the Add Health data.

By contrast, it is possible to create better models of who enters into what type of relationships and what factors drive them to make those choices by leveraging machine learning techniques. This approach involves predicting who marries, cohabitates, or remains single from as large an array of predictors as possible using a rich, longitudinal dataset (i.e., the Add Health data set). Given its utility, others have recently begun to apply machine learning in this fashion to large-scale, longitudinal data, including Add Health to examine outcomes besides marriage. For example, Esposito et al. (2017) used machine learning to understand the effect contact with the criminal justice system has on health. Additionally, Hill et al. (2019) used machine learning to identify individuals who are at-risk for suicide.

In contrast to these studies, the present study utilized a machine learning approach to

analyze lifestyle choice using multiclass classification to look at marriage, cohabitation, and singlehood simultaneously within a single model, using a more extensive set of variables than those used in prior work. The purpose of this analysis was to understand better which variables are the most important and to uncover any that have been overlooked by prior research in the marriage, cohabitation, and singlehood domains. In other words, the novelty of this study is that it: 1) applies machine learning specifically to the problem of predicting lifestyle choice; 2) uses a multiclass outcome consisting of marriage, cohabitation, and singlehood to explore these outcomes simultaneously; and 3) incorporate as large a set of variables as possible to uncover nuanced relationships between the predictors relative to the lifestyle choice outcome.

That said, it is important to note the current study focuses on outcomes occurring at Wave IV, at which point individuals have reached young adulthood (i.e., ages 24 to 32). Many life-changing events can occur after the age of 32, and therefore the full scope of the results of these analyses are limited by the dataset under evaluation. Analyzing additional waves or conducting further studies that are able to examine individuals across their lifespan may yield more conclusive results that capture important changes that may occur later in life.

In summary, this study aims to answer an overarching question about lifestyle choice: What are the most critical variables that predict individual lifestyle choice, and how do these predictors change over time? Understanding who will eventually end up in a non-marital or non-cohabiting relationship will enable psychologists to identify at-risk individuals, understand their backgrounds and behaviors, and ideally uncover ways to encourage more beneficial outcomes in the future.

Literature Review

Today, the number of ways individuals choose to have relationships, including their

living arrangements, is increasing. Commonly in 20th-century American culture, couples would enter into a marriage and then live together. Marriage describes situations in which individuals choose to enter a legally recognized relationship with one another. Recently, many western countries have undergone shifts in the structure of relationships as two additional relationship states have become more significant to demographic research: cohabitation and singlehood. Cohabitation shares many qualities of marriage (e.g., intimacy, presumed monogamy, shared residency, division of household labor, shared parenting). However, cohabitation does not offer the same legal distinction as marriage. In contrast to these two dyadic states, singlehood is a phenomenon that describes an individual who consciously chooses to remain single into old age.

Various factors can act as early predictors for determining whether an individual will choose marriage, cohabitation, or singlehood. Such determining factors help researchers characterize our developmental models of an individual's predisposition to any one of the three described states. Effective developmental models can better inform the basis and timing of practical intervention and provide useful targets when identifying avenues for research.

The following review covers 1) definitional issues that describe each state; 2) empirical predictors that help explain how and why individuals choose or are disposed to one of the three described states; 3) the behavioral findings relevant to an individual's health and wellbeing associated with each state; 4) the current demographic findings and interpretations that can be used to contextualize the three states descriptively; 5) and in a separate section, an examination of the three states comparatively to highlight essential similarities and distinctions.

Marriage

Marriage comes with many benefits, including companionship, improved mental and physical health, and personal family benefits such as improved childhood outcomes and

increased effectiveness in parenting (Thomas & Sawhill, 2002; Lichter, 2001). However, marriage as a social construct is undergoing tremendous changes, and it is becoming increasingly difficult to understand the predictive factors of marriage, including the decision not to marry. Some of the complex factors involved are individual preferences, life experiences, personal and societal beliefs, and changing socioeconomic norms. Moreover, many studies on the factors of predicting marriage have been conducted as post-marital, rather than pre-marital studies. The lack of pre-marital data may have led to skewed or inaccurate results that were then used to answer the questions related to which factors are involved in predicting marriage. This literature review takes into account multiple sources that indicated interrelated factors for predicting marriage, and of these, the three primary factors are 1) personal circumstances (primarily from one's youth); 2) personal beliefs; and 3) socioeconomic status.

Overview

Both internal and external factors predict whether an individual will choose to get married (Shmerling, 2016). The three primary predicting factors are personal circumstances, beliefs, and socioeconomic status. DeLap (2000) and Tumin (2016) illustrated how personal circumstances influence a person's relationship choices. For example, people who had childhood-onset disabilities and those who had alcoholic parents tended to shy away from marriage more than their counterparts later in life (DeLap, 2000; Tumin, 2006). Both of these sources concluded that people who had childhood struggles tended to experience mental illness and poor social development later in life, resulting in an inability to form intimate relationships. These long-term effects of a troubled childhood subject individuals to relationship strain increased divorced rates and lower marriage rates (Whisman, 2006).

Two previous studies that examined the impacts of positive childhood experiences found

that people who had stable childhood households were more likely to get married (Kefalas et al., 2011; Larson & Olson, 2004). These two studies also found that personal beliefs also play a role in predicting whether someone will get married: Specifically, Kefalas et al. (2011) and Larson and Olson (2004) found that placing value on the institution of marriage plays a significant role in predicting a person's decision to marry. Individuals with these beliefs often come from affluent, metropolitan, and religious families. Furthermore, these same individuals also tend to believe that marriage is a milestone of life that must be achieved (Kefalas et al., 2011). Similarly, Larson and Olson (2004) found that religious individuals, as well as people who believe they share a spiritual connection with their partner, see marriage as a requirement for fulfilling their religious or spiritual duties.

In addition, economics is also an important aspect of predicting marriage. Cherlin et al. (2016) and Rackin Gibson-Davis (2017) both concluded that the higher the household income for any given couple, the greater the likelihood that the couple would marry. The researchers found that finances play a significant role in marriage decisions, and individuals in higher income brackets tend to place a higher significance and greater emphasis on the social construct of being married (Cherlin et al., 2016).

Health and Marriage

The rationale for investigating the predictive factors of marriage may be explained by analyzing the benefits of marriage. At a conference for the British Cardiovascular Society, researchers presented their findings on marriage and myocardial infarction (MI) mortality rates (Hayes et al., 2016). The researchers conducted a regression analysis on 25,287 individuals newly diagnosed with MI, with control variables including age sex and gender. The mean age of the participants was roughly 70 years of age, of which approximately 64% were male, and 80%

were Caucasian. The mean length of stay (LOS) was 7.0 days. Based on their findings, the researchers concluded that marital status and mortality rates were linked. Individuals were identified using eight categories: single, married, divorced, common law living, unmarried, separated, unknown. While most distinctions do not need to be defined (e.g., single, divorced, separated), the researchers did not clarify the difference between single and unmarried. Unmarried and married individuals had between 2.12 and 2.66 days shorter LOS' than singles, while widowed and separated individuals had an additional 1.82 to 2.66 days longer LOS' than singles. The researchers attributed the support, or lack thereof, for a patient depending upon their relationship status as the underlying cause. Such a claim is supported by a bevy of evidence demonstrating many marriage-related benefits.

One study from the late 1990s indicated that marriage positively impacted and promoted healthy behaviors, including maintaining a well-balanced diet and increasing physical activity (Steinberg-Schone, 1998). Such behaviors are linked to improvements in both physical and mental health. More recently, a study of 740 adults analyzed stress levels among three groups - married couples, never-married couples, and previously married individuals - to determine the participants' overall health (Chin et al., 2017). However, participants younger than 21 were excluded from the study due to the extremely low rate of marriage, leaving a total of 572 participants. In this study, background characteristics were controlled for, and stress levels between the groups were compared. Three studies were conducted in total, whereby participants conducted screenings via phone and received a physical examination from The results indicated that married couples had lower levels of cortisol compared to the other two groups. Researchers attributed this to health-related behaviors that are fostered in the interpersonal relationship

between married individuals.

Another study conducted in 2016 highlighted some health benefits associated with married couples: longer lifespan, fewer strokes, lower rates of depression, fewer occurrences of cancer, and increased likelihood of surviving cancer and major surgeries (Shmerling, 2016). The researchers concluded that, in general, these health benefits were a result of the reinforcement of positive lifestyle changes that influenced long-term health (e.g., sleep, diet, and exercise).

Alternatively, marriage has also been shown to have negative impacts on overall health if the relationship exhibits frequent marital strife (Liverpool, 2018). These adverse effects include poor mental health, high levels of stress and anxiety, and decreased immune system effectiveness. However, these problems are not permanent. Researchers have found that troubled couples who get divorced could potentially alleviate many of these negative health effects associated with a lower quality marriage (Kendler, 1987).

Demographics

Marriage rates have been declining, and being unmarried or being engaged in an alternative relationship choice has become more socially acceptable than in the past (Fry, 2012). Between the years 2000 and 2018, marriage rates were at their highest in 2001 at around 2.3 million new marriages, and by 2009, marriage rates had reached a low of nearly 2.1 million (CDC, 2020). Of the 2.2 million marriages in 2017. However, Romero (2017) found that roughly 490,000 same-sex marriages in the US in 2017 accounted for approximately 20% of marriages that year - an effect attributable to legal changes due to Obergefell v. Hodges (Murray, 2016). While marriage rates are declining across all demographics, the poorly educated are experiencing

the greatest decline (Lundberg et al., 2016).

Conversely to the decline in marriage rates, population levels grew from roughly 280 million in the year 2000 to 325 million in the year 2020 (a 15% increase; Pollard et al., 2020). These figures, in conjunction with the marriage statistics table, make it clear that the marriage rate has been decreasing despite an increase in the population.

One possible explanation for the decline in marriage rates since 2000 is the change in our societal norms (Gubernskaya, 2010). Nowadays, people choose to adopt alternative relationships, such as cohabitation and singlehood. As opposed to the expectations of "marriage naturalists," for many others, marriage is only one of the choices an individual can make for their future (Kefalas et al., 2011, "Abstract," p. 27).

As a Result of Personal Circumstances

An individual's personal life plays a vital role in determining the likelihood of getting married later in life (DeLap, 2000; Kefalas et al., 2011; Turnin, 2016). DeLap (2000), Kefalas et al. (2011), and Turnin (2016) presented independent predictive factors of marriage, such as unfortunate childhood situations and the general quality of one's upbringing. DeLap (2000) originally posited that there is a relationship between having alcoholic parents as a child and later deciding not to get married. While there was no statistically significant correlation between this hypothesis and the collected data, a multitude of personal, mental, and social development issues were present as a result of having alcoholic parents. It is possible that these traits later deterred the individual from marriage.

Such traits were similarly identified in a 16-year study analyzing the correlation between childhood disability and marriage in over 560,000 individuals (Tumin, 2016). The results indicated that children diagnosed with early-onset disabilities demonstrated a reluctance to get

married in adulthood. One possible explanation of this phenomenon is that disability creates several social challenges and functional barriers to forming a relationship. From the standpoint of psychology and behavioral health, such challenges include social anxiety, depression, and an inability to connect and develop intimacy with others. Tumin's (2016) developmental psychological research identified how detrimental childhood experiences are negatively correlated with an individual's inclination to marriage (i.e., increased negative childhood experiences were negatively correlated with marriage).

Furthermore, individuals in one study reported making choices based on their life goals and marriage values. These choices and marriage probability and early life marriage-related goals were positively correlated (Kefalas et al., 2011). The majority of participants reported similar personal circumstances. Specifically, they often endorsed having experienced stability during childhood and having positive role models. However, such situations are not the only predictors of someone's decision to marry: Broader reasons such as religious beliefs, or beliefs in certain social constructs, can also be predictors.

As a Result of Personal Beliefs

Religion and social norms play a large part in deciding to get married (Kefalas et al., 2011). Based on an individual's reported marriage goals, Kefalas et al. attempted to determine if background characteristics could be used to make descriptive and predictive distinctions between those who are and are not likely to marry. The study concluded that there were two distinct groups: "marriage naturalists" and "marriage planners" (Kefalas et al., 2011, "Abstract," p. 27). The naturalists were more likely to be from a religious background, live in rural areas, and be less likely to cohabitate before marriage. In contrast, the planners were more likely to live in metropolitan areas, have lower religiosity than the naturalists, cohabitate before marriage, and

choose to delay childbirth.

Numerous studies have shown a positive correlation between religious or spiritual beliefs and marriage (Larson & Olson, 2004). Larson and Olson conducted a meta-analysis that reviewed and analyzed 15 sources that consistently indicated that a couple's consensus on their religious or spiritual beliefs is indicative of their marriage prospects. In other words, Larson and Olson's research predicted that an individual who identifies as belonging to a religion is much more likely to get married compared to their non-religious counterparts. Whether or not this is a result of commonality in religious belief or for nonreligious reasons is disputed; however, researchers agree that religion may predict marriage since couples who shared the same religion were more likely to share a common demographic background, and therefore share a conventional belief system that further promotes the probability of marriage. Marks (2008) identified eight influential factors among religious people - three of which agree with Larson and Olson's (2004) findings: the practice of marital fidelity, which provides support for co-belief in a religious worldview; pro-marriage beliefs that promote marriage; and mutual faith in God that acts as marital guidance and support.

As a Result of Socioeconomics

The cost of marriage is difficult to quantify and is often a barrier for many who wish to get married (Edin & Reed, 2005). Moreover, the consistently rising costs of living (e.g., expenses of raising a child, utilities, housing costs, and groceries) serve as consistent barriers to marriage and are thereby predictors of a couple's decision to marry. To be specific, couples who have higher incomes are more likely to get married since more wealth means the couple can afford the associated costs of marriage.

Cherlin et al. (2016) completed a study that focused on relationship formation.

Individuals were divided based on their income and then divided again into two groups: medium-to-low incomes and medium-to-high incomes. The rates of non-marital births and marriage were then compared for both groups. The researchers concluded that income predicted a couple's decision to get married to an extent. Specifically, non-marital births were significantly higher in the low-to-medium income group, whereas marriage rates were significantly higher in the medium-to-high income group. These differences were attributed to the cost of marriage and societal norms associated with higher income brackets (i.e., people in the higher-income groups were more likely to value getting married before having children).

Other studies have found data consistent with Cherlin et al. 's (2016) findings. Rackin and Gibson-Davis (2017) conducted a qualitative analysis of 69 lower-income individuals. Within this group, a distinct trend emerged: child-rearing did not hold the same societal expectations as marriage. The researchers concluded that while there are not necessarily extraneous costs associated with marriage, there were subconscious expectations of financial benchmarks that must be met before a couple can get married. Cherlin et al. (2016) and Rackin and Gibson-Davis' (2017) research firmly support the conclusion that income or wealth is predictive of a person's likelihood of getting married; most specifically, these studies indicated lower-income individuals get married less often than higher-income individuals.

Summary

Many predictive factors could influence an individual's decision to marry. These factors may be based on personal circumstances, personal beliefs, or economic factors. While each factor has its weight in marriage prediction, economic and educational factors may be the most predictive. Although marriage is a widely accepted social norm, the rate of marriage is decreasing each year, despite the potential positive health benefits associated with marriage.

However, alternatives to marriage (e.g., cohabitation and singlehood) have been increasing in popularity, especially among the poorly educated.

Cohabitation

In some Western societies, marriage used to be the only legitimate social construct deemed appropriate for two adults who desire to have a monogamous intimate relationship with one another (Modell, 1980), and cohabitation prior to 1970 was minimal (Lundberg et al., 2016). Numerous authors have illustrated that cohabitation occurs before marriage in many cases (Brown et al., 2006). However, this has changed as cohabitation (i.e., a relationship in which intimate partners live with one another but are not planning to get married) has become more culturally viable and more common, lifelong option for couples, although the driving force behind this change has multiple possible causes (Rhoades et al., 2012). Similar to marriage, multiple factors predict whether or not an individual will get married later in life. Rhoades et al. found at least three key factors that are reliable in such a prediction: personal circumstances, gender, and contextual factors.

Overview

Cohabitation is a living arrangement wherein two individuals in a relationship decide to live in the same household together without being married. Although this phenomenon has gained significant popularity in recent years, its growth can be traced to the late 1960s: Cohabitation grew from 0.1% in 1968 to 9.4% in 2018 (Gurrentz, 2018).

Three factors consistently predict cohabitation: personal beliefs, age and gender, and personal reasons or circumstances. Solely based on these three factors, various interrelationships are present, which will be discussed in a later section of this review. Using these interrelationships, Rhoades et al. (2012) and Brein et al. (2006) correctly predicted that people

who possess and share more traditional and religious beliefs about marriage and relationships would not cohabit during their life.

On the other hand, people who share beliefs in the benefits of living with someone before marriage often cohabitate (Rhoades et al., 2009; Tang et al., 2014). These researchers illustrated that one of the main reasons people cohabit is the desire to spend more time with their significant other. Both Rhoades et al. (2012) and Tang et al. (2014) argued that this desire is a consistent predictor of cohabitation.

The final predicting factors are age and gender. Certain demographics are more likely to cohabit than others (Brown et al., 2006; Stepler, 2017). While age and gender are not stand-alone factors, they play an essential role in predicting cohabitation when incorporated with multiple factors. For example, men make up the majority of older cohabitors and are more likely to cohabit with younger women (Brown et al., 2006).

Health and Cohabitation

Existing studies on companionship have mostly focused on the benefits of marriage; however, many health benefits are reflected in both marriage and cohabitation. Many cohabiting people perceive their relationship as equally significant as marriage (Perelli-Harris et al., 2017). The cohabitating subjects in Perelli-Haris et al. 's study were found to receive many of the same health benefits as married couples. One such benefit was the encouragement each partner provided the other to monitor lifestyle choices and engage in healthy behaviors. These behaviors help ensure the longevity of their relationship and live a long and healthy life.

Zimmerman and Easterlin (2006) conducted a study to determine life satisfaction across various lifestyle choices (i.e., marriage, cohabitation, and divorce). The researchers took into account life satisfaction as it naturally varies across "sex, age, income, education, health,

employment, and religiosity" (p. 7) before conducting their analyses. The data was collected from the German Socioeconomic Panel using Waves 1 through 21, which spanned across the years 1984 to 2004. Time-variant covariates were also included. To conduct the analysis, the researchers asked, "How satisfied are you with your life, all things considered?" (p. 8) and then used a centering technique to account for the aforementioned variables. The results indicated that successful relationships (i.e., those which do not end in divorce or separation) were positively correlated with improved wellbeing, with no distinction between marriage and cohabitation. In other words, being in a committed relationship (i.e., marriage or cohabitation) has measurable, positive benefits on health and general wellbeing.

Other researchers have found similar results to those of Zimmerman and Easterlin (2006): Rettner (2012) found that marriage and cohabitation are equally beneficial in terms of overall health. It is important to note that Perelli-Harris et al. (2012) pointed out that married couples may receive one health benefit that couples who cohabit do not: shared health insurance policies.

Demographics

Demographic profiles play a large part in predicting whether or not individuals will cohabitate (Stepler, 2017; Gurrentz, 2018; Gurrentz, 2019). Brown et al. (2006) analyzed how age acted as a sufficient predictor of cohabitation. It is important to note that Brown et al. defined cohabitation as a heterosexual couple living together in an intimate relationship while remaining unmarried. Data was collected from two sources (19,727 respondents) - the 2000 Census and the 1998 Health and Retirement Study from 1981 - and multinomial logistic regression models were used to determine the relationship between cohabitation and remarriage among adults aged 55 and older. Within this sample population, gender played a significant role in the data. Women largely outnumber men in terms of being unmarried, while approximately

60% of cohabitors over the age of 60 are men.

US Census data showed an increase in cohabitation among men ages 50 and over by 75% from 2007 to 2016 (Stepler, 2017). It can thus be inferred that an individual within this demographic would be more likely to cohabitate later in life. Men who have never married or cohabited by their 40s will most likely remain single for the rest of their life, further illustrating that cohabitation and singlehood (discussed later) have become relevant and prevalent alternatives to marriage (Fineman, 1981).

Another study conducted by Manning et al. (2019) identified trends to account for Stepler's (2017) previously estimated 24% increase in cohabitation occurrences among young people in the US. Manning et al. (2019) sampled 2,700 young women between the ages of 18 and 24 and found that young adult women are more likely to cohabit compared to young adult men and middle-aged individuals, regardless of sex. However, the high rates of cohabitation among young adult women are not indicative of pure cohabitation. Specifically, these women cohabit as a precursor to marriage.

As previously discussed, the popularity of alternatives to marriage has been increasing since the early 2000s (Brown et al., 2006; Manning et al., 2019). The number of cohabiting adults in the US increased from 2007 to 2016 by 29%, and this increase included a growing number of individuals over the age of 50 (Stepler, 2017). Furthermore, there have been increases of 24% and 20% in age groups 18 to 34 and 35 to 49, respectively. This increase appears to correlate positively with an increase in divorce rates.

Census data released in 2018 focused on two age brackets: 18 to 24 and 25 to 34 (Gurrentz, 2018). In 1968, 0.1% of individuals ages 18 to 24 and 0.2% of individuals ages 25 to 34 were cohabiting. By 2018, those percentages increased to 9.4% and 14.8%, respectively.

Conversely, the percentage of married individuals within the same age brackets significantly decreased from 1968 to 2018. Married individuals ages 18 to 24 had a significant decrease from 39.2% to 7.3%. Similarly, individuals 25 to 34 who were married decreased from 81.5% to 40.3%. Gurrentz suggested these changes may have been the result of the Great Recession and a lack of financial security. The author also linked socioeconomic status to marriage rates: lower socioeconomic couples' marriage rates decreased much more rapidly than higher socioeconomic couples.

As a Result of Cultural Background and Beliefs

Personal beliefs in this context are limited to the individual's perceived belief in marriage as an institution and the individual's religious beliefs. The first of these is discussed thoroughly by Rhoades et al. (2012), whose study analyzed the causal relationship between the belief in the constitution of marriage and cohabitation. Rhoades et al. 's original hypothesis focused on the correlation between a belief in marriage and cohabitation, predicting that a lack of belief would indicate a decreased likelihood to marry. However, the study revealed that an individual's lack of belief in marriage was not predictive of cohabitation.

In a 2002 study that analyzed cohabitation, Brein et al. (2006) found numerous predictive factors by measuring a coefficient concerning the utility gained by the involved individuals. Religion played a significant role in predicting cohabitation: Catholics were far less likely to cohabitate than non-Catholics. Based on the analysis by Brein et al., personal beliefs, particularly religious beliefs, can play an essential role in predicting whether or not people will cohabit.

Both Brein et al. 's (2006) and Rhoades et al. 's (2012) findings can be further explained by analyzing individuals' beliefs. For example, some individuals maintain a traditional way of thinking and therefore participate in traditional relationships (e.g., marriage). Traditional

lifestyles are more common among religious individuals and do not support unconventional relationships or divorce (Leifbroer & Rijke, 2019). This social norm among religious communities may explain why religious individuals report lower cohabitation rates than their nonreligious peers who have fewer traditional beliefs (Thornton et al., 1992).

Other studies highlight the importance of shared beliefs within a relationship and how such commonalities can influence a couple's likelihood of cohabitation or marriage. More specifically, individuals with more shared beliefs were more likely to cohabitate or marry (Hohmann-Marriot, 2006).

Personal Preferences. Other predictive factors are not as easily quantifiable as either demographics or religion, such as the personal feelings that people tend to possess regarding relationships (Rhoades et al., 2009). Cohabitation preceded more than 50% of marriages according to a paper from 2006 by Brown et al., and cohabitation accounted for nearly half of the relationships of the US adults ages 18 to 65 in 2019 (Gurrentz, 2019). Many of these individuals indicated similar reasons for engaging in cohabitation before marriage (Rhoades et al., 2009).

In a 2009 study of 240 individuals, Rhoades et al. (2009) found that a majority of couples were found to have the same reasons for wanting to be in a relationship with one another. Each individual was given a questionnaire focused on why they wanted to engage in cohabitation prior to marriage. They were then asked to rank each reason on a scale according to whether they agreed or disagreed with each qualitative statement. After analyzing the questionnaires, Rhoades et al. found two primary reasons individuals choose cohabitation: increasing time spent with their partner and increasing intimacy. Tang et al. (2014) also found that individuals choose to cohabitate to increase the amount of time they spent with their partner before marriage. Tang et al. 's study found that this desire to spend more time together consistently predicted more

positive relationships, in which both commitment and overall satisfaction were increased. While these studies offer no statistical predictions, they offer insight into why people choose to cohabitate and support previous predictions, such as the importance of common goals and shared beliefs.

Summary

These data show a trend in the increased popularity and acceptance of cohabitation as an alternative to marriage. Cohabitation offers a legitimate and suitable option for those who do not believe in or do not wish to follow the social institution of marriage, and also serves as a beneficial mediator between singlehood and marriage. Many factors can predict whether or not an individual will choose to cohabit, and many benefits and reasons support those predictions. Demographics, beliefs, religion, and personal opinions all play a role in determining the likelihood of an individual cohabiting in the immediate future or later in life. Regardless, the literature suggests that cohabitation is becoming an increasingly powerful and popular social institution in which people can experience companionship and relationships.

Singlehood

A person's lifestyle choice may not be based solely upon personal preference alone. A multitude of decisions, life circumstances, and other contributing factors may affect an individual's likelihood to be in a relationship. These contributing factors are important to consider when analyzing singlehood. Although it is challenging to ascertain a direct correlation as to what causes an individual to choose singlehood, some factors may be used as predictors. This section will review five unique sources that all provide unique and interrelated factors that may be used to predict singlehood. These factors include personal circumstances, personal

decisions, and gender.

Overview

The number of single-person households is on the rise in many countries, although not all individuals living in single households go on to become single across their lifespan (Fokkema & Liefbroer, 2008). Perpetual singlehood occurs when an individual consciously chooses never to marry and remain single into old age.

Existing literature suggests that at least three major factors can predict singlehood: personal decisions, life circumstances, and gender (Allen, 1994; Bellani et al., 2017; Band-Winterstein & Manchik-Rimon, 2014). According to these authors, personal decisions that individuals make early in life can determine whether they will be a singleton for a variety of reasons, including whether or not they choose to attain higher education, pursue a lifelong career, or be successful.

Life circumstances - such as having to take care of parents or family, childhood illness or cancer, or poor quality of parental marriage - play critical roles in predicting singlehood. For example, Gurney et al. (2009) conducted a study using over 10,000 people in the Childhood Cancer Survivor Study cohort and found that nearly 50% were single, including many who were over 35, suggesting lasting consequences attributable to childhood illness. In addition, Cunningham and Thornton (2006) argued that parents' marital quality has a transmissible impact on children.

Band-Winterstein and Manchik-Rimon (2014) found that women in Israel were more likely to remain single into old age, and Bellani et al. (2017) found similar patterns in other countries: patterns that can be attributed to womens' need to make choices between careers, education, independence, or having a family. On the other hand, Janson (2009) showed that men

who had adverse childhood experiences were more likely to remain single. These authors illustrated three common predicting factors to singlehood, which were also mirrored by cohabitation and marriage, as described below.

Health and Singlehood

In the early 2000s, multiple studies - including those of DePaulo and Morris (2006) and Hacker (2001) - illustrated the benefits of marriage relating to access or affordability of automobile and health insurance, and how single people are often excluded from those benefits. It should be obvious that the lack of access to healthcare can have tremendous adverse effects on an individual's health, thereby suggesting there might be a correlation between singlehood and health status in countries where universal healthcare access is not guaranteed.

Pudrovska et al. (2006) also analyzed the relationship between singlehood and health. The strain an individual feels as a result of being single, termed "single strain," was defined as an indicator of chronic stress, which can lead to physiological and psychological illnesses, such as headaches, irritability, insomnia, digestive issues, depression, low self-esteem, decreased libido, and anxiety. Researchers used a group of 530 older adults and analyzed individuals' cortisol levels and determined which group was affected the most by single strain. The study concluded that never-married white women shouldered the most significant proportion of single strain, indicating that singlehood does cause adverse health effects.

Some research suggests there are a few benefits of singlehood, including the ability to maintain a broader array of individual friendships (Sarkisian and Gretel, 2016). These singletons were also able to travel and engage in a variety of activities without the constraints typically associated with marriage and cohabitation. The freedom and independence associated with singlehood can lead to some benefits, including stress reduction, improved social interactions,

and a reduction in other stress-related health complications. However, singletons also experience decreased sexual gratification and may have a more difficult time becoming an integral part of the community (Laplant, 2016).

Demographics

As previously stated above, the demographics of lifestyle choice is changing, and marriage rates are declining, and over time, more individuals are choosing cohabitation and singlehood over marriage. Unlike marriage and cohabitation, there is a relative paucity of research on the numbers of people who choose to live as singletons.

The number of unmarried, single Americans was estimated by the CDC (2017), and the findings were published in an online article in 2017. In the study, federal-level census data from 2017 was used to measure the number of singletons living in the US. The analysis suggested that around 110.6 million individuals over the age of 18 were single. Among that group, around 70.2 million were unmarried and had never been married before. Finally, when examining older adults (those over the age of 65), the researchers found that the number of older individuals who were never married was close to 19.2 million. Although singlehood cannot be accurately inferred from this data, it suggests that a large number of people are single when they are younger, but some individuals remain single for the majority of their lives.

Gender is perhaps the most straightforward predictor of singlehood, but one of the most difficult to explain. Various analytical tools, such as retrospective analytics, as well as multi-level analyses, have been used to predict whether or not men or women are statistically more likely to remain single later in life. Bellani et al. (2017) conducted a multi-level analysis on a sample size of nearly 85,000 individuals from countries within the European Union and concluded that men are statistically more likely to remain single in old age. Specifically, roughly

7.3% of men and 4.2% of women remained single into old age, with more fine-grained variation across countries.

Researchers have identified some specific factors that contribute to why women are more likely than men to be singletons. Cultures in which society has replaced traditional gender roles with more flexible expectations see higher rates of single women than men (Bellani et al., 2017). Bellani et al. presented two different conclusions as to which gender more conclusively predicted singlehood later in life, both based on the level of gender egalitarianism present in the culture. Gender egalitarianism is positively correlated with a higher likelihood of women choosing singlehood than their male counterparts.

Finally, Canadian census data up to 2016 showed a drastic rise in single-person households over the past several decades (The Daily, 2017). One possible explanation behind this continued increase is the change in sociocultural norms, which emphasizes values related to independence. This explanation is supported by vastly changing gender roles, with an increasing number of women choosing to remain single, be more independent, and pursue lifelong careers (Pastor, 2008). In short, there is a shifting trend in the composition of demographics of lifestyle choice.

As a Result of Personal Circumstances

As with marriage and cohabitation, life circumstances play a crucial role as early predictors for singlehood (Allen, 1994). For example, an individual's familial responsibilities (a lifelong circumstance) have been shown to act as a reliable predictor of singlehood. Other life circumstances that influence a person's lifestyle choice are societal roles, cultural norms, and childhood experiences.

Life circumstances are often dictated by social roles and constraints set by the society in

which an individual lives (Huston, 2000). For example, caring for elders is a common societal norm in many cultures with the expectation that adult children will act as their parents' caregivers when their parents enter old age (Lin & Wu, 2019). Because such situations are a sociocultural phenomenon, the individual does not have control over their personal life (Band-Winterstein & Manchik-Rimon, 2014). Regardless of familial obligations, the responsibilities that fall upon an individual have been described to act as predictors of, and explain why these people choose singlehood.

Additionally, research suggests that parental marriage quality perceived by an individual during childhood is significantly related to singlehood in adult life (Cunningham & Thornton, 2006; Kefalas et al., 2011). Cunningham and Thornton's (2006) research found that parental discord was positively correlated with children being dissuaded from marriage, an attribute correlated with a higher level of adult singlehood.

Another researcher analyzed the correlation between childhood experience and lifestyle choice later in life. Janson et al. (2009) conducted a study using approximately 9,000 individuals from the Childhood Cancer Survivor Study and found that 46.4% of the participants who were diagnosed with cancer never married, and 90% of these participants were living single, concluding that the adverse experiences of childhood cancer strongly influence an individual's lifestyle choice later in life. Tumin (2016) also found that childhood disabilities and illnesses often force individuals to delay romantic partnerships and that they are more likely to remain single.

As a Result of Personal Decisions

One of the most significant decisions for an individual is whether they will pursue an advanced degree. This decision can then significantly impact a person's decision to engage in an

intimate relationship. Men with higher education levels experience higher rates of singlehood than women (Bellani et al., 2017). Bellani et al. showed a 5.5% probability of singlehood among women who obtained higher levels of education compared to 3.5% with medium or low levels of education. By contrast, the finding for men was reversed: up to 6% probability of singlehood among men who obtained higher levels of education, and 11% for those whose education was low. Hill (2020) conducted interviews with 47 Ph.D. students to study singlehood, gender expectations, and education. All 47 participants were unmarried and did not have children and ranged from 22 to 36 years old. Using dual-pass coding methods, Hill analyzed the data and found three trends: perceptions of cultural norms (e.g., starting a family); anxieties surrounding work-life balance; and plans on establishing a future work-life balance. Based on this study, Hill concluded that there are two primary types of young singletons: those who are single and intend to remain as such and those who are delaying romantic relationships to pursue their education or career. More often than men, women expressed concerns over balancing work and family life, while men did not typically indicate any desire to slow down their pursuits for a family.

Moreover, career choice also acts as a predictor of singlehood. Yoshida (2011) found that one of the main reasons individuals choose to remain single is their job requirements, meaning the desire to be successful in life greatly influences their lifestyle choice. Although career choice and pursuing higher education do not directly result in singlehood, they act as early predictors. All of these sources provide substantial evidence to support predictions of singlehood and illustrate that conscious life choices can lead to singlehood.

Summary

Various qualitative and quantitative factors, including sociological, cultural, familial, and

personal, can serve as accurate predictors of singlehood. The reviewed literature provided both individual and collective predictors, many of which concluded that gender is a baseline for both statistical and qualitative analyses. Gender was then broken down into cognitive and non-cognitive decisions, further predicting an individual's decision to choose singlehood over other choices, such as marriage or cohabitation. The analyses also illustrated many comparisons can be drawn between marriage and singlehood, since the same factors that predict marriage align with those that predict singlehood. However, cohabitation's predicting factors should also be taken into consideration and how they affect the factor relationship between marriage and singlehood. Regardless, it is evident that no single reason or decision leads people into a life of singlehood.

Summary

Marriage, cohabitation, and singlehood are all different types of living situations, and the lifestyle choices a person makes depends upon their beliefs, personal circumstances, demographics, health benefits, and a host of other factors. Individuals who yearn for companionship may lean towards cohabitation or marriage, while those who desire freedom and independence may choose singlehood. While unique in their own right, each of these situations has many commonalities that exist between them. Further analysis will be needed to determine which factors uniquely predict a specific lifestyle choice, which relates to more than one, and so forth. Machine learning analysis provides an avenue for this type of discovery.

Predicting Outcomes

The overarching goal of this project is to understand the psychological and physiological influences relevant to relationship outcomes across an individual's lifespan. This study will accomplish this task by using machine learning techniques to analyze longitudinal population

survey data from the Add Health dataset and identify models that can accurately predict relationship outcomes in young adulthood (ages 24 through 32) at Wave IV. Machine learning is ideally suited to these analyses due to the Add Health dataset's complexity: Specifically, this method is capable of identifying the most relevant predictors among the thousands within Add Health while simultaneously offering a way to examine the temporal effects that influence these predictors. This researcher will use predictive machine learning to improve our understanding of individuals who choose to marry, cohabitate, or remain single while considering the highly individual, multimodal pathways that influence these relationship outcomes across the lifespan.

To conduct these analyses, a variety of machine learning models were employed to discover the factors that predict lifestyle choice. By evaluating multiple models, this study hopes to shed light on an important set of questions: 1) How well can marriage, cohabitation, and singlehood be predicted within the Add Health dataset?; 2) Do certain topic constructs in Add Health such as substance use or personality influence these predictions more so than others among the widest range of predictors possible?; 3) Are there variables from earlier in the lifespan that can accurately predict outcomes that occur later in life up to young adulthood?

The general approach that is used is consistent with the life course method in developmental psychology and sociology (Baltes et al., 1980), which examines how variations in theoretically and empirically valuable domains, such as social relationships, employment, and education influence, determine one another through intricate patterns of interaction, which are time-dependent. For example, King et al. (2004) examined how rates of childhood externalizing and internalizing can be used to infer substance use (or abuse) at a later period in the life course, such as in adolescence and early adulthood. The analytic framework of these studies involved using multidimensional longitudinal data to identify complex systems that work to produce

essential outcomes empirically; In this respect, the Add Health dataset is ideal for this analysis of lifestyle choice-related outcomes. Moreover, in contrast to other research that has attempted to answer similar predictive questions in longitudinal datasets regarding topics of marriage and cohabitation, this study uses machine learning and cross-validation to produce more robust results than past efforts while incorporating a broader range of predictive variables (e.g., Black et al., 2001; Fincham & Beach, 2010).

Finally, this project harnesses an underutilized approach to variable identification/selection in the context of psychological research. The use of machine learning in the life-course analysis is not new, but not yet widespread (Billari et al., 2006). Existing social science research on marriage and cohabitation has primarily focused on simple statistical models and/or regression techniques; therefore, the incorporation of machine learning could lead to heretofore unexplored, novel findings. Ultimately, this project aims to confirm existing relationships between previously studied variables included in the Add Health dataset and identify new relationships that have yet to be explored. Finally, this study is intended to serve as a template for conducting such analyses with other potential outcome variables of interest (e.g., body mass index [BMI] or intelligence).

Method

The following sections describe the use of multiclass classification machine learning in the Add Health dataset. The Add Health dataset's key aspects are described by providing: 1) a general overview of the dataset; 2) details regarding the dependent variable (DV) related to lifestyle choice prediction; and 3) details regarding the independent variables (IVs). Secondly, the study design and analytic strategy are discussed, including 1) data preparation and cleaning,

2) the machine learning model types used in this study, and 3) the machine learning analysis workflow.

Dataset and Participants

The Add Health dataset is a longitudinal study of individuals from the US who were followed from adolescence into adulthood across multiple study waves. The sample population was collected from 80 high schools and 52 middle schools that were non-randomly selected from US schools, stratified based on county, urbanicity, school size, school type, and ethnicity (Harris et al., 2009). Add Health was created in 1994 to fulfill a government mandate to study adolescent health and was funded by project grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) and twenty-three supporting organizations.

The dataset currently consists of five waves. However, this researcher conducted the analysis using Waves I through IV, as shown in Table 1 below. This was due to the late-release of Wave V and the researcher's inability to gain access to Wave V data. In the first collection of Wave I, 90,118 students from the target schools in grades 7 through 12, completed an in-school administered survey. From this group, a subset of 20,745 adolescents was selected for an in-home survey/interview.

Waves II through V continued to track these participants over two decades. In Wave II, 14,738 participants from grades 8 through 12 were surveyed, 15,197 individuals, aged between 18 and 26, along with 1,507 of their romantic partners in Wave III, and 15,701 individuals between the ages of 24 and 32 in Wave IV. Finally, by the time of Wave V, these individuals will be between the ages of 32 and 42, although data for this Wave had not yet been released while this study was conducted.

Across the waves, demographic and biographical information related to each individual

are surveyed and recorded in the study data. However, access to these data is limited to prevent deductive disclosure risk. Table 1 below summarizes the Add Health population across the available waves (i.e., Waves I through IV).

Table 1
Add Health Wave Structure

	Wave I	Wave II	Wave III	Wave IV
Sample Size	20,745	14,738	15,197	15,701
Age Range	12 - 20	12 - 20	18 - 26	24 - 32
Age Description	Childhood	Adolescence	Emerging Adulthood	Young Adulthood
Years of Data Collection	1994 - 1995	1996	2001 - 2002	2008 - 2009

Each wave is composed of items and questionnaires intended to meet the changing research goals associated with studying participants across an individual's life. Waves I and II were intended to broadly measure social relationships and community connection, whereas Waves III and IV were designed mainly to examine health-related behaviors.

More specifically, the content of the surveys and interview questionnaires at Wave I and II covered the following topics: health status, health-facility utilization, nutrition, race and ethnicity, peer networks, decision-making processes, family composition, and dynamics, educational aspirations and expectations, employment experience, the formation of romantic partnerships, sexual partnerships, substance use, and criminal activities. Wave III added new surveys beyond those previously included, and these covered the following topics: relationships, marriages, childbearing, educational, and workforce events. Wave IV also added several new surveys with additional topics pertinent to adulthood, such as education and employment

transitions, changes in financial resources, and more.

Although the in-home surveys used in each of the successive waves contained many similar surveys and survey items, as mentioned earlier, differences exist from wave to wave. Most of these differences can be attributed to the slightly different empirical, methodological, and scientific goals adopted by Add Health researchers as the project moved forward, along with age-specific life events in the cohort across the waves.

Defining the Analyses

The broader purpose of this analysis is to predict which individuals within the Add Health dataset have either married, cohabited, or remained singletons by Wave IV using a machine learning approach. Since the outcome is known a priori, this represents a supervised - as opposed to unsupervised - machine learning problem. Moreover, because the outcome has three discrete levels, this represents a multiclass classification problem.

Given the complex, multi-wave nature of Add Health, it is possible to predict lifestyle choice in various ways by focusing on predictors from different waves. For example, one might want to know if data collected between Waves I and III can accurately predict lifestyle choice that occurs later on at Wave IV. Alternatively, one might choose to create a more inclusive, descriptive model and use predictors across all available waves (i.e., at Waves I through IV, inclusive) to predict within Wave IV. Ultimately, this study conducts both analyses, with the former intended to be more focused on prediction and the latter on description when it comes to lifestyle choices in young adulthood.

Identifying the Dependent Variable (DV)

Based on the Relationships subsection, it is possible to determine class membership at Wave IV using two items: H4TR1 and H4TR2. The first item, H4TR1, states: “How many

persons have you ever married? Be sure to include your current spouse if you are married now" (Add Health Codebook Explorer, 2017), and H4TR2 states: "How many romantic or sexual partners have you ever lived with for one month or more? By 'lived with,' we mean that neither of you kept a separate residence while you were living together". Using the answers to these two items, it is possible to derive the class membership of each participant in the following manner: Any participant who answered H4TR1 indicating one or more marriages has class membership in the marriage set. Any participant who is never married as per H4TR1 and also answers H4TR2 indicating that they have cohabitated has class membership in the cohabitation set. Any participant who has no membership in either group as per responses to H4TR1 and H4TR2 has membership in the singlehood set. Ultimately, this process produced a single, three-class dependent variable (DV) consisting of the following classes: 1) Married at Least Once; 2) Pure Cohabitation; and 3) True Singleton.

Identifying the Independent Variables (IVs)

The independent variables for this analysis are drawn from the surveys in the Add Health dataset across Waves I through IV. The effort was taken to incorporate as many surveys and items as possible, although some had to be removed out of necessity to produce meaningful models. For example, the two items mentioned earlier to create the DV were removed (i.e., H4TR1 and H4TR2) along with similar items that directly asked participants to describe their lifestyle choice (e.g., H3MR1: "How many times have you been married?") or were otherwise dependent upon it or closely related to it.

For example, consider item H3EC57 in the Economics and Personal Future section of Wave III: "What do you think are the chances that the following will happen to you? You will be divorced by age 35." This item was problematic since divorce directly related to marriage, and

some of the response options were overly informative (e.g., “This has already happened”). Nevertheless, other items from the H3EC section were retained since they did not directly connect to the DV. For instance, item H3EC14 was retained: “Do you have an email account?” - interestingly, this item ends up being important to one of the predictive models presented later.

More broadly, because certain sections focused exclusively on marriage, cohabitation, or lifestyle choice, they were excluded outright from specific waves (e.g., Marriage/Co-habitation History and Attitudes in Wave III; Relationships in Detail in Waves III and IV). A more detailed description of the removed sections and items is provided in Appendix A. In the end, considerable time was spent going through the items in the Wave item index files, the Add Health Codebook Explorer (2017) topics listing, and the results of preliminary models to ensure confounding items were accounted for and removed prior to conducting the analyses.

Study Design and Analytic Approach

Table 2 below shows an overview of the machine learning workflow used for the analyses, which is divided into four broad stages (i.e., Stages I through IV) that cover a total of six steps (e.g., Steps 0 through 5). As shown below, Stage I covers the basic steps necessary for Data Cleaning and Preprocessing (Step 0); Stage II describes the Machine Learning Models that were used in this analysis (Step 1); Stage III covers Model Evaluation and Comparison using multiple metrics (Steps 2 and 3); and Stage IV is a discussion of Model and Variable Interpretation (Steps 4 and 5).

All of the analyses in this study were conducted in R (R Core Team, 2020) with a heavy reliance on the caret package (Kuhn, 2008), which offers a streamlined way to prepare data, run multiple machine learning models, and evaluate model results. It is worth noting that despite using caret, individual model types each require their own, distinct packages as discussed later

when detailing the various model types employed in this analysis. The caret package unifies the syntax for running these models across packages and provides a variety of helpful functions and packages for data preparation, model comparison, visualization, etc.

Table 2
The Machine Learning Workflow

Step	0	1	2	3	4	5
Stage	I: Data Cleaning and Preprocessing	II: Machine Learning Models	III: Model Evaluation and Comparison		IV: Model and Variable Interpretation	
Objective	Prepare data for analysis, including variable selection, partitioning, etc.	Select machine learning models and specify starting parameters	Fit models and find best parameters within each model type	Select the best overall model based on multiple evaluation metrics	Identify the variables of importance within the best model	Follow-up analyses of the top identified predictors

Data Cleaning

The opening sections of this paper detailed why Add Health is a suitable dataset for machine learning and described the particular DV and IVs that will be used to conduct a classification analysis. However, in terms of data preparation and cleaning, several key steps had to be taken to select and prepare/clean the data for analysis.

The first of these relates to the wave structure of Add Health: namely, the number of participants changes from wave to wave (e.g., see Table 1). Because the DV is based on individuals' lifestyle choice at Wave IV, the initial sample is limited to the 15,701 participants who were included in the data at this point. Nevertheless, this is still a massive sample compared to traditional longitudinal analyses (e.g., for comparison, the Boston Couples Study mentioned earlier only included a total of 231 couples; Peplau et al., 1993).

The second concerns the particulars of the DV: a small number of individuals did not have valid responses on the marriage and cohabitation items used in its construction (i.e., H4TR1

and H4TR2 as discussed earlier). For instance, a small number of people refused to supply an answer to one or both of these items. In total, this meant that of the 15,701 initial instances, the DV could only be computed for 15,662 of them (i.e., a total of 39 were excluded by necessity). Table 3 below provides an approximation[1] of the number of these individuals who reported having ever married at each wave based on relevant variables (i.e., H1GI15: "Have you ever been married?"; H2GI3: "Since {MOLI}, did you get married?" [where MOLI indicates month and year of the prior Wave I interview]; H2GI5: "What is your current marital status?"; H3MR1: "How many times have you been married?"; and H4TR1: "How many persons have you ever married? Be sure to include your current spouse if you are married now"). As shown in the table, by Wave IV, nearly 50% of individuals in the Add Health dataset have married at least once.

Table 3
Number Married Across Waves I to IV

Wave	Number Married
I	62
II	138
III	2,495
IV	7,797

The third and most complicated sub-step relates to how items are coded in Add Health. Add Health is a rich and complex dataset, containing over 8,207 columns after merging across Waves I through IV. Across the various waves, the surveys include a mixture of continuous (e.g., H4HS3: "Over the past 12 months, how many months did you have health insurance?"), categorical (e.g., H4CJ17: "Have you ever spent time in a jail, prison, juvenile detention center or other correctional facility?", and Likert variables (H4RE1: "What is your present religion?).

[1] Due to cases of occasional missingness and some minor inconsistencies in the Add Health data across waves, these numbers should be treated as a close estimate rather than an exact count.

Having this rich variety of different types of variables is ideal for evaluating a wide range of machine learning models (Guyon & Elisseeff, 2003), and ideally the variables should be modeled with respect to the underlying type of data being represented (e.g., continuous data should not be discretized). However, achieving this is complicated using the Add Health data since various codes are included across most items to indicate circumstances such as refused responses (e.g., represented by response code 98 in H4RE1 while the standard responses were coded from 1 to 9) or legitimately skipped responses (e.g., represented by response code 7 in the binary H4CJ17 item mentioned above). These distinct response codes can present a problem when working with continuous data, since they do not represent actual responses but rather special cases. As such, it was necessary to filter these responses out by treating them as missing data (i.e., NAs), which was done using a variety of techniques including making use of the Add Health Public-Use Data for generating updated item coding (via the SPSS data files which contain actual labels unlike the CSV files which only supply the numeric codes) and filtering on various response codes that indicated special-case responses (e.g., response labels such as "invalid," "refused,", "legitimate skip," "over limit," etc.). Ultimately this meant that only those variables that appeared in the Public-Use Data were used to conduct my analysis, despite having access to additional items in the Restricted-Use Data.

In addition, a small number of items (< 50) were removed from the analyses for representing complicated factors with more than 20 levels. Examples of these items include H1GI12, "In what country were you born?", and H3MN2, "How is this person [H3MN1] related to you? If there has been more than one person, describe the most influential." Moreover, an additional ~30 items were removed due to additional complexities related to item coding (e.g., H2GH43, which involved time data: "During the summer, what time do you usually go to bed on

week nights?"; H3LM10, which involved job codes: "What did you do in that first job?") or because they related to the survey design (e.g., BREAK_Q: "Breakoff questions asked").

Finally, as described in the preceding section, an additional set of items were removed from the analysis because they were tied directly to the dependent variable. This included 17 code book sections involving household rosters, relationship information, pregnancies, and marriage/cohabitation history and a handful of additional items that are further described above in the section on Identifying the IVs and elaborated upon in Appendix A. In sum, after removing these various items, a total of 7,855 variables spanning Waves I through IV remained in the data from the initial 8,207, including the DV.

Data Preprocessing

Prior to fitting machine learning models, additional steps must be taken to set up the data for the analysis. Specifically, at a minimum, it is necessary to partition the data, consider class imbalance, check for missing data, and filter out certain problematic variables that can interfere with the modeling procedure. This type of filtering falls under a broad, prerequisite step in fitting most machine learning models known as data preprocessing. In the following subsections, an overview of the main preprocessing steps in this analysis is provided. As a reminder, the data contained 15,662 rows (individuals) and 7,855 columns at the start of this stage, including the dependent variable of lifestyle choice with three distinct classes: Married at Least Once, Pure Cohabitation, and Singleton.

Data Partitioning. Data partitioning is a requirement to fit and evaluate machine learning models (Kuhn, 2013). In the case of my analysis, prior to fitting any models or preprocessing the data , I randomly divided the dataset into an 80/20 training/testing split, which is a common partitioning ratio recommended in the machine learning literature (e.g., Suthahatan,

2016). In other words, I used 80% of the data to fit the machine learning models and reserved 20% for model performance validation in each analysis. The key aspect of this training/test split is that the 20% "holdout" sample will represent entirely new data that a trained model will never have seen before; as such, these data cannot influence model training and thus provide an unbiased way of evaluating model performance in a novel, randomly pre-selected data sample. In other words, validation on the holdout sample provides a way of showing that a model with high performance in the training data can replicate this high performance in the secondary validation dataset.

Class Imbalance. Much of the discussion surrounding the data thus far has focused on issues that relate to the IVs. However, there are important decisions to be made regarding the DV that influence how well the machine learning models perform. The first of these issues relates to the concept of class imbalance. Because the proposed studies examine a multiclass outcome, it is necessary to ensure an equal number of individuals represent each outcome category (or "class") within the DV: The reason this is important is because machine learning models can produce biased results when the classes are highly disproportionate, complicating model evaluation and interpretation of predictor significance and model performance (Kuhn, 2013). Downsampling is one solution to this problem that involves keeping all cases of the minority class (i.e., the one with the smallest proportion of cases) and a random subsample of the majority classes (e.g., in a sample of 100 people, assuming 50 people in the training data were coded as As, 30 as Bs, and 20 as Cs [i.e., 50/30/20 = 100 total], then a downsampled training dataset in this case would consist of a random sample of 20 As, Bs, and Cs [20/20/20 = 60 total]).

Given the large size of the sample, downsampling was appropriate in this analysis to avoid potential problems and to achieve high prediction accuracy that was class-agnostic,

consistent with machine learning best practices (Kuhn, 2013). Moreover, despite downsampling, each of the classes in the DV were well represented in the subsequent analyses, with thousands of cases used from training in both models: specific details about the number of cases are provided later in this paper.

Dummy Variables. While certain types of machine learning models such as C5.0 can handle factor data natively (Quinlan, 1993), many cannot handle data in this format and expect the factors columns to be converted to a series of dummy columns. For example, if a factor variable called IV100 contains response options A, B, and C, then it would be necessary to convert this variable to a set of binary columns that represent this coding differently (e.g., one way is to use three columns - IV100_A, IV100_B, and IV100_C - that contains 0s and 1s to indicate whether the response on IV was that specific letter [or not]; in other words, if a row value for IV100 contained a B, then IV100_A, IV100_B, and IV100_C would be coded as 0, 1, 0, respectively). This researcher used the dummyVars function in caret to take care of this problem (Kuhn, 2013).

Low Variance Predictors. All items which have zero or minimal variance must be removed from the dataset prior to fitting models, since these items cause issues with model convergence—an issue that also affects basic regression models. Moreover because these items have very low or no variance, they consequently offer little to no predictive value when it comes to the DV.

To understand why low variance items are not useful, consider a column in the dataset that records whether or not a person was born within the past century with the following dummy coding: 1 = yes; 0 = no. Everyone in the Add Health dataset would have a value of 1 for this item, and as such, this item would have a mean of 1 for all individuals regardless of whether or

not they have ever married (i.e., relative to the DV). Moreover, the variance of this item would also be 0 for individuals regardless of their status with regard to the DV. An item with equal means across groups and no variance obviously provides no statistical value. For this reason, items that had no variance were removed prior to fitting models in my analysis, consistent with best practices when it comes to machine learning (e.g., Kuhn, 2013).

Missing Values. Some machine learning algorithms cannot be used when the data have missing values: the models will simply refuse to fit. Although one option is to remove participants who happen to have missing data (i.e., listwise deletion), this brute force approach can often be suboptimal since it may lead to removing many participants' data for even a single missing observation. In addition, this type of removal may also introduce systematic and potentially undetected bias with regard to the effective data sample and thus harm the subsequent analysis, particularly when data are not missing at random. Fortunately, imputation provides a better alternative: rather than removing data with missing values, imputation provides a way to replace missing values with reasonable alternatives. Many methods exist for imputing missing data, including common simple approaches such as mean/median substitution, nearest neighbors (e.g., Beretta & Santaniello, 2016), multiple imputation (e.g., Azur et al., 2011), etc. Median imputation was used during model training to handle missing data in my analysis for models that required it (i.e., random forest and SVM). In addition, variables from the training data that contained 80% or more missing values were removed (see the Appendix A for more details).

Machine Learning Models

As a reminder, one of the reasons machine learning was used for this analysis is because the Add Health dataset is rich, consisting of thousands of variables from dozens of different surveys, over ten thousand participants, and a multi-wave longitudinal structure. This type of rich

"big" data is well-suited to machine learning analysis.

In addition, a complex topic such as marriage and cohabitation may be characterized by the interaction of many different variables that show different patterns or relationships within distinct groups of people. Simple, straightforward regression models tend not to work well when the data involve thousands of variables and complex interactions between them, and existing research has shown that these models can provide misleading estimates when using them for prediction due to a high likelihood of overfit (e.g., Black et al., 2001).

Moreover, when conducting these types of analyses related to problems such as how social systems interact with personality, studies have often relied on correlational methods, simple regression, and/or a large degree of simplification and lack of sufficient control variables (e.g., see Amato (2015), which attempts to justify the latter concern). Because machine learning methods are designed to incorporate thousands of variables in a complex manner that can uncover nuanced patterns within the data, this analytic approach offers a more robust alternative for studying these types of complex topics than more simple techniques such as regression while simultaneously avoiding issues related to overfit such as those discussed by Black et al. (2001), who raised this concern in the context of predicting marriage outcomes using longitudinal data and recommended cross-validation as a potential solution to the problem.

Cross-Validation. Cross-validation is an extremely common technique used in machine learning to validate model results during the training procedure and avoid concerns such as data overfitting (Kohavi, 1995). Moreover, cross-validation is also often relied on to find optimal hyperparameters, which are sample-independent, model-specific settings that may affect how well certain models perform. In light of this, cross-validation was used in this analysis to accomplish several important goals: 1) to avoid potential overfitting concerns; 2) to optimize

hyperparameter selection; and 3) to increase the likelihood that model results are replicable across multiple training iterations, thereby reducing the likelihood that good model fit would be attributable to random chance.

The specific type of cross-validation employed in this analysis is k-fold cross-validation, which involves splitting the training data into several partitions, k in total, such that some k - 1 partitions are used for training, while one is reserved for evaluating model fit as part of an iterative process. For a basic example of k-fold cross-validation, suppose the initial dataset to be modeled consists of 2,000 rows, with 1,500 used for training (75%) and 500 for test (25%). While the latter 25% of the data is reserved and cannot be used for model training, the 75% training data can be further split into three equally-sized thirds to conduct a simple, 3-fold cross-validation, resulting in sets A, B, and C with 500 rows each. Using these three sets, a model can be trained using data from sets A and B (1,000 rows total), while data from set C are reserved to evaluate how well that model fits "novel" data in C that was not directly used for model training. Afterwards, to get a sense of training variability, the process can be iterated across the folds: specifically, in the 3-fold case, models trained using sets B and C can also be evaluated on A, and sets A and C can be evaluated on B. Thus, a 3-fold cross-validation procedure ultimately involves a total of three training iterations, with a greater number of iterations possible by choosing higher values of k. Again, a key advantage of this approach is that the cross-validation procedure helps reduce the likelihood of overfitting during training by seeing if the results are generalizable across the various random folds of the training dataset. For this analysis, 5-fold cross-validation will be used, which is a common fold setting used for model building (e.g., Kuhn, 2013).

As mentioned earlier, another practical advantage of using cross-validation is that it can

be used to aid in the selection of model hyperparameter settings. Specifically, cross-validation can be used in conjunction with a grid search, which is a process that involves trying different combinations of multiple model settings drawn from a large, combinatorial grid (e.g., see Bergstra and Bengio (2012) for a discussion of grid search and more complex methods such as random search). Working through the setting grid while iterating across cross-validation training folds provides an empirical way of determining how different parameter settings influence model fitting results while also getting a sense of model fit stability. A variety of different hyperparameters specific to each model type were tested as part of this analysis (e.g., see Appendix B for a plot of settings used in the best models).

Finally, as implied in the preceding paragraph, in addition to using cross-validation for model training, multiple machine learning model types were evaluated to find the overall best model for predicting the multiclass outcomes of marriage, cohabitation and singlehood. Evaluating multiple models is valuable not only to increase the likelihood of finding the model that is most predictive, but also to ensure that the analysis is robust and not simply the byproduct of the particular model type employed.

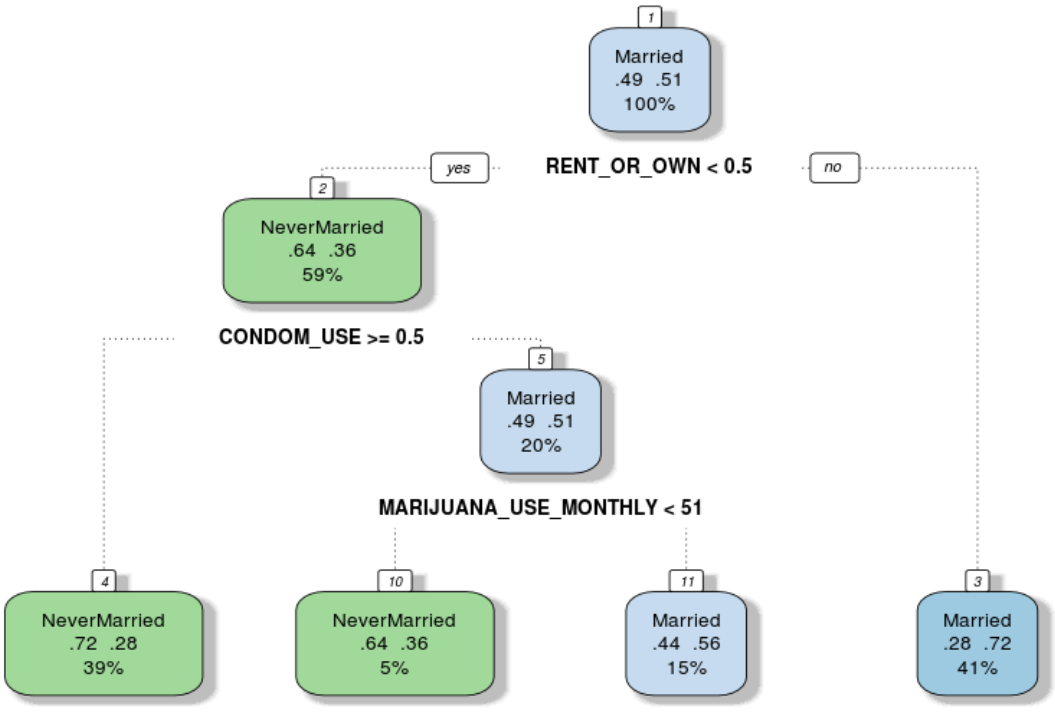
The following sections describe the different types of machine learning models used in this analysis one-by-one. These models fall into two broader categories: decision tree models and support vector machines (SVMs). The first discussion will review decision tree models broadly and then turn to three model types that are all related to decision trees: gradient boosted models, random forests, and C5.0. Afterwards, a brief introduction to SVMs is provided, which collectively represent the second major family of models used in this analysis. Again, all of these models were run using caret, which integrates models from a variety of individual packages into

a unified syntax.

Decision Tree Models. The decision tree is a type of predictive model for the simultaneous analysis of the relationship between multiple variables relative to a categorical or continuous outcome (Quinlan, 1986). Although decision tree learning is the basis of a wide range of machine learning models (e.g., gradient boosted models, random forests, etc.), it is helpful to understand the basic concept of what decision trees entail. To accomplish this, examine an example derived from a relatively basic type of decision tree model known as classification and regression trees (CART; Breiman et al., 1984), which can be generated in R using the rpart package (Therneau, 2017), which is supported by caret.

Figure 1

CART: Married vs. NeverMarried



Note. A basic illustrative example of the output of a simple classification and regression decision tree model (CART) for a potential binary outcome of Married vs. NeverMarried.

A decision tree can be described as a visual "tree-like" system made up of nodes, branches and end nodes, each of which when put together represent possible decisions and corresponding outcomes. Figure 1 above is an example of the visual output from a very simple decision tree model that predicts a binary outcome - in this case married vs. unmarried - from an assortment of predictors including, rental vs. home ownership, condom usage vs. non-condom usage, and a measure of monthly marijuana usage at Wave IV. A core element of a decision tree relates to "splits," which involve decision paths based on predictors used within the overarching model.

Although the visual inspection of decision trees can be informative, depending on the number of variables involved, it may not ultimately be helpful when many variables are incorporated in the model due to the higher degree of potential complexity, although this can be controlled to some degree by varying a complexity parameter (cp) in the model. What is more important is the underlying representation, which can be used as the basis for making predictions for new cases that were not used to develop the model (i.e., novel test data). Moreover, the tree can also be used to determine which predictors are very important relative to the outcome and which are not using measures of variable importance.

Gradient Boosted Models. The prior section focused on CART as a basic example of decision tree models. Many other types of models exist within this family, and one that has become popular involves a technique known as gradient tree boosting: a process of creating a single tree model from a collection of weak tree models, which collectively form an ensemble (Ridgeway, 1999).

To provide a bit more detail, a gradient boosted model (GBM) starts by generating a

series of weak tree models referred to as “stumps.” A stump is characterized by a single split rather than multiple splits, meaning that a stump is composed of a single primary node, two leaves, and two terminal nodes. Several of these stumps are then added together by the GBM sequentially to create an additive model, which is optimized based on a loss function in an effort to reduce errors, minimize bias, and constrain model complexity to avoid overfit (Freund & Schapire, 1999). This general process can be extended to incorporate trees of greater complexity than just stumps, in addition to controlling various other parameters such as the number of iterations when forming tree ensembles.

The GBM model is available in caret using the gbm package (Greenwell et al., 2019). Hyperparameters that were evaluated included the number of trees in the model, the number of observations in each node, shrinkage (or learning rate), and interaction depth, which - as the name implies - allows for testing variable interactions inside the trees.

Random Forest. Random forest models are similar to gradient boosted models in that they also involve creating an ensemble of decision trees (Ho, 1995). However, unlike gradient tree boosting which combines a number of weak tree models, random forests are built up of fully grown trees. When it comes to classification, each fully grown tree (e.g., a CART from the decision tree section) within the random forest creates a prediction; by finding which prediction is the most common, the random forest ultimately decides which class a test case is mostly likely to represent (e.g., if a random forest model contains 500 trees, a single prediction may result in 400 0s and 100 1s; in this case, the final prediction will be a 0). The problem of using these complex, unpruned trees in random forests is that it could result in a high degree of overfit, although hyperparameters allow for some degree of control to avoid this issue.

Random forests were evaluated in my analysis because they are commonly used to solve

machine learning problems with a high level of success (e.g., Statnikov et al., 2008). An efficient implementation of random forest models is provided in the ranger package in R (Wright & Zeigler, 2017). Hyperparameters that were evaluated for random forests included the number of variables and instances used when splitting at nodes as well as different splitting rules (e.g., Cutler et al., 2007).

C5.0. The C5.0 model - which is itself an extension of another model known as C4.5 (Quinlan, 1993) - is another type of model within the overarching decision tree family. C5.0s offer several unique features not available in the other models discussed so far. Two important features of C5.0 include: (1) the ability to create rulesets in addition to decision trees for binary classification; and (2) the potential to further reduce model complexity through a process known as winnowing. C5.0 models are available in the R package C50 (Kuhn et al., 2015).

Regarding the former, in simple terms, a rule is just a conditional if-then statement (e.g., IF Sex is 'male' THEN MarriedAtLeastOnce is FALSE). Rules can be combined together into a mutually compatible system or "ruleset" (e.g., IF Sex is 'male' AND IF Age > 30 AND WantsChildren is 'yes' THEN MarriedAtLeastOnce is TRUE). While fundamentally similar to decision trees, rulesets have an advantage of simplifying interpretability since they lack a tree structure; moreover, rule-based models are often less complex than tree based models and can also be more accurate in some cases (RuleQuest, 2017).

A second useful feature of C5.0 is winnowing, a method which reduces the number of variables incorporated into a predictive model to make it simpler, more generalizable, and more interpretable. More specifically, in cases when there are many predictors available for building a model, oftentimes several variables are particularly useful for building a good model and while others are less so: winnowing weights predictors based on how useful they are and removes

unimportant variables prior to model fitting to decrease complexity and thereby improve the final model's generalizability. A downside of winnowing is that it can take a lot of time to perform the underlying procedure, but the upside is that it can lead to stronger models (RuleQuest, 2017).

C5.0 models were incorporated in this analysis using the C50 package in R (Kuhn & Quinlan, 2020). For hyperparameters, both tree and rules models along with varying the winnowing setting were evaluated.

Support Vector Machines. Support vector machines (SVMs) represent an entirely different class of models from decision trees. Nevertheless, SVMs are similar to decision trees in that both are often used for supervised binary classification tasks (e.g., Fan et al., 2008). Similar to many of the other models discussed in this section, SVMs can also be used for regression problems as well, although this is not the focus on my research.

In the case of binary classification, SVMs attempt to find the optimal separation between two classes or entities within a dataset using one or more hyperplanes. For linear SVMs, a hyperplane is defined as the line that maximally separates cases in one class from cases in the other(s): a process known as a maximum margin (Burges, 1998; Kotsiantis et al., 2007).

For this analysis, a linear SVM from the e1017 R package was used (Meyer et al., 2019) since this variation allows for computing class probabilities while working with caret. A variety of cost settings for the model during training were also tested.

Analysis I

For the first predictive model using the Add Health data, the dependent variable (DV) was lifestyle choice observed at Wave IV, and the independent variables (IVs) were drawn from Waves I through III. As discussed in the section entitled Defining the Analysis, the data used for this analysis were filtered such that only those individuals who had not reported a marriage up

through Wave III were included in the modeling process. Of the 15,662 cases in the initial dataset, this ultimately left 10,499 eligible for Analysis I.

As previously described, five different types of machine learning models were evaluated: decision trees (rpart), gradient boosted models (gbm), random forests (ranger), C5.0 (c50), and support vector machine (svm). The data were partitioned into 80% training (8,400 rows) and 20% testing (2,099 rows), and after data cleaning and preprocessing, the total dimensions of the downsampled training set were 5,262 rows with 1,754 cases each of Married at Least Once, Pure Cohabitation, and Pure Singleton in the multi-class DV and 8,458 IVs.[2] A five-fold cross-validation technique was used to train all the various models on identical sub-samples, and various hyperparameter settings were evaluated for each model type using a grid search approach with accuracy as the evaluation metric. The best version of each model was ultimately evaluated using the non-downsampled 20% test data, which consisted of 795 Married at Least Once, 866 Pure Cohabitation, and 438 Pure Singletons (for a total of 2,099 as noted earlier).

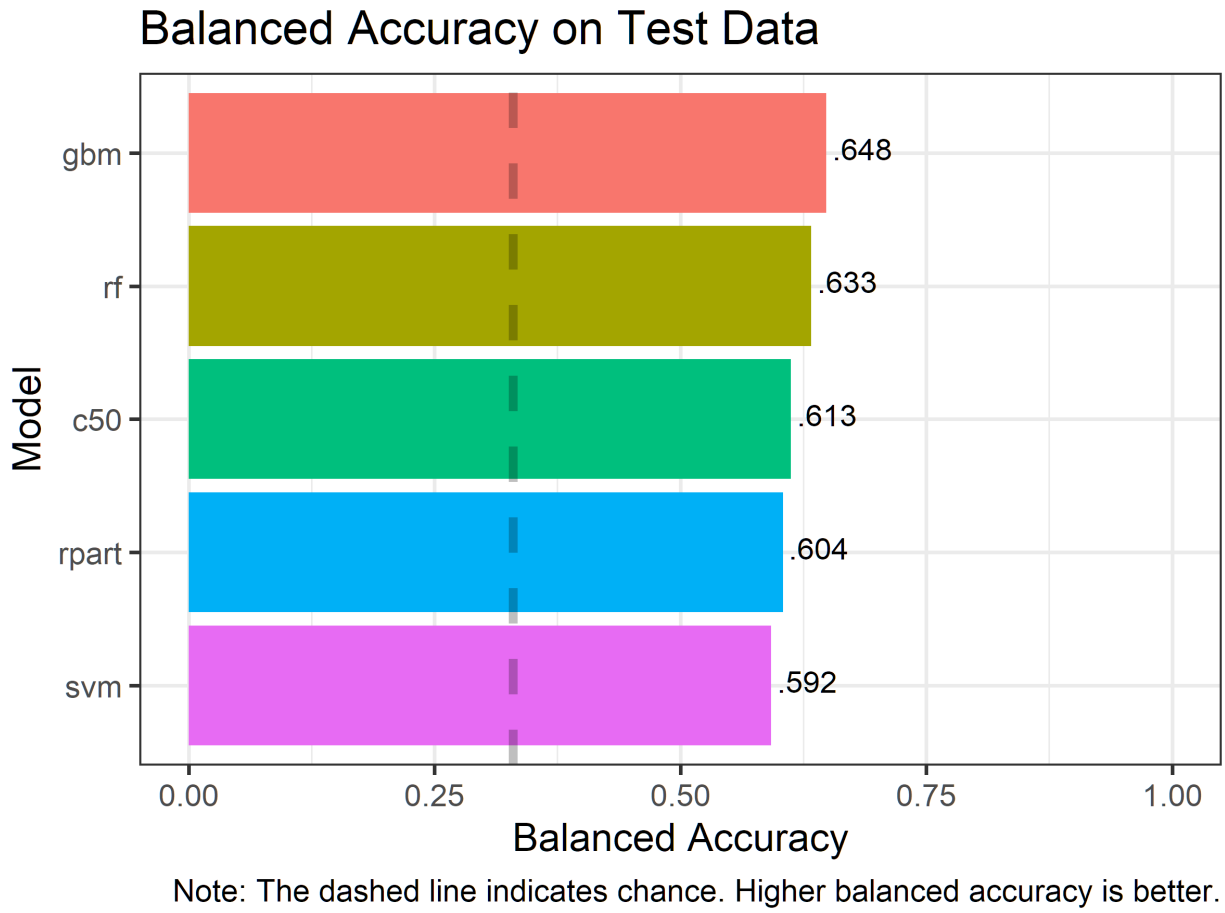
Model Comparison

Figure 2 below shows the overall performance of these best models on the test data using macro balanced accuracy (i.e., the average accuracy across the three classes in the DV). Performance for all models was well above chance (i.e., one in three for a three-class DV, or .333), and the highest performing model is shown at the top of the figure, which was the gradient boosted model (gbm). Overall accuracy for this model was .648, meaning it was accurate on approximately two out of three test cases when predicting lifestyle choice at Wave IV using predictor variables from Waves I through III.

[2] The section on data cleaning noted there were 7,855 variables across Waves I through IV, but the number of variables in Waves I through III was 6,952. Moreover, some of these variables were removed due to reasons described earlier such as no variance, a high number of missing data points, etc., and some were dummied, which increased the tally. After accounting for all data preparation, 8,458 is the number of IVs used in Analysis I.

Figure 2

A comparison of balanced accuracy in the training data for Analysis I.



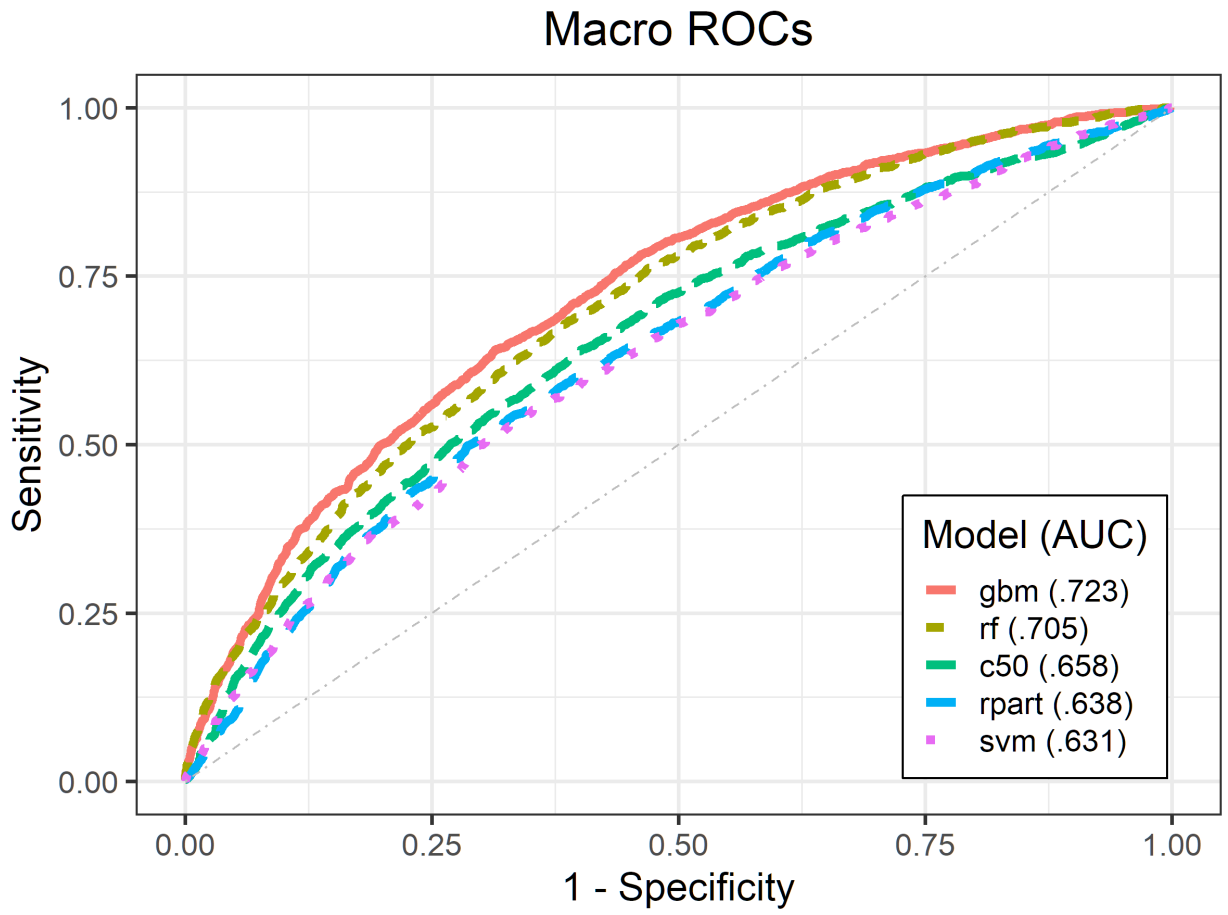
Note. The gradient boosted model (GBM) performed best overall and well above chance. The dashed line indicates chance. Higher balanced accuracy is better.

In addition to balanced accuracy, the models were compared using macro averaged receiver operating characteristic (ROC) curves and area under the curve (AUC) scores. As the name implies, AUC scores represent the area under the ROC curves, with higher values closer to one indicating better performance (i.e., ROC curves that move closer to the top left of the figure are ideal). As shown in Figure 3 below, similar to balanced accuracy, the gbm model performed the best overall compared to its competitors with an AUC of .723.

Figure 3

Macro averaged receiver operating curves and AUC scores for each model in the test data for

Analysis I.



Given the gbm model performed best in terms of both balanced accuracy and AUC, it was selected as the final model for the analysis and for subsequent variable interpretation. For additional details regarding model performance in the training data cross-validation and for details regarding the influence of hyperparameters in the gbm model, see Appendix B.

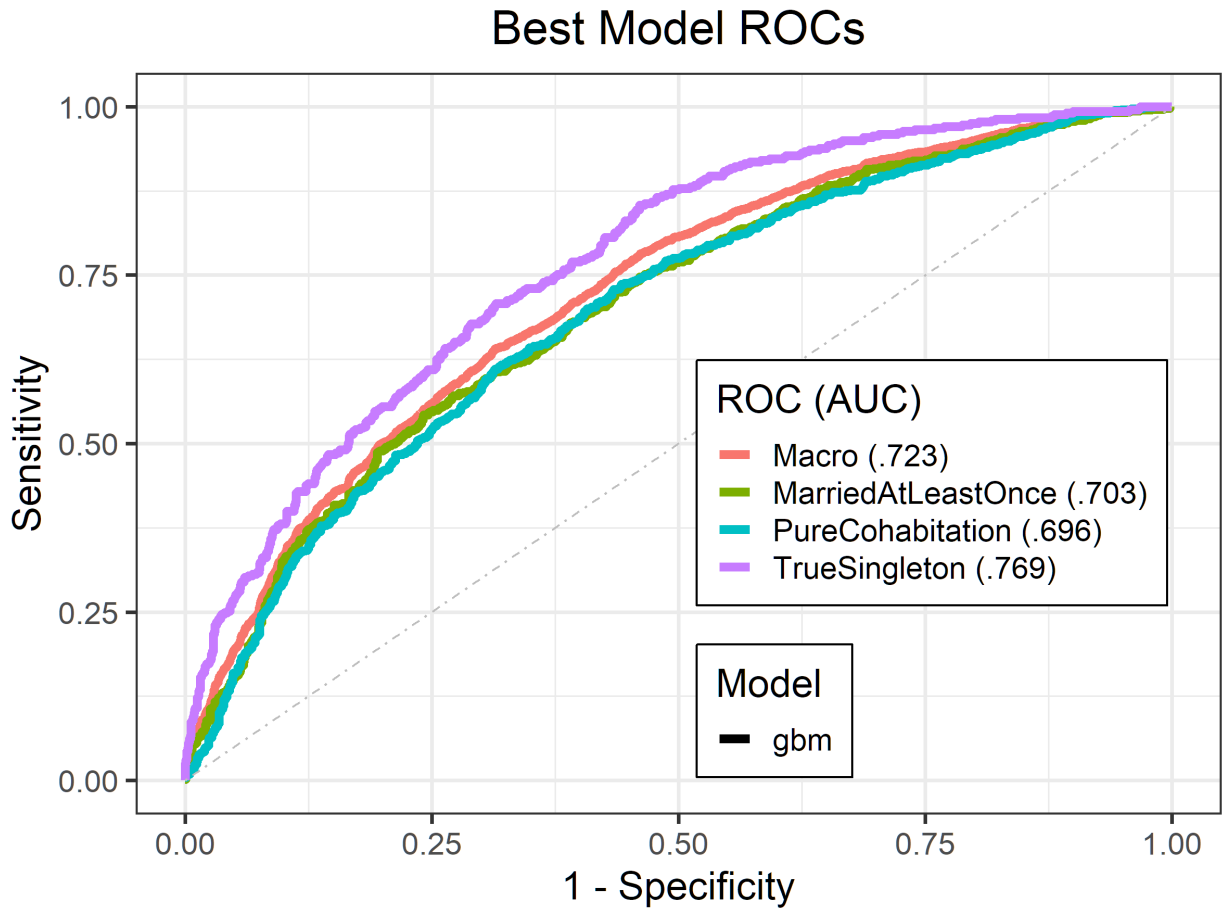
Investigating the Best Model

Figure 4 below shows a set of individual ROC curves by class for the gbm best model, as well as the same macro average ROC shown in Figure 3 (red line). Based on the AUC scores and associated ROC curves, the model performed best on average when predicting True Singleton status (purple line). By comparison, the ROC curves for Married At Least Once (green line) and

Pure Cohabitation (blue line) exhibited lower performance compared to the other two outcomes and pulled down the macro average.

Figure 4

The macro ROC curve for the best gbm model in Analysis I and the constituent ROC curves for each class within the dependent variable.



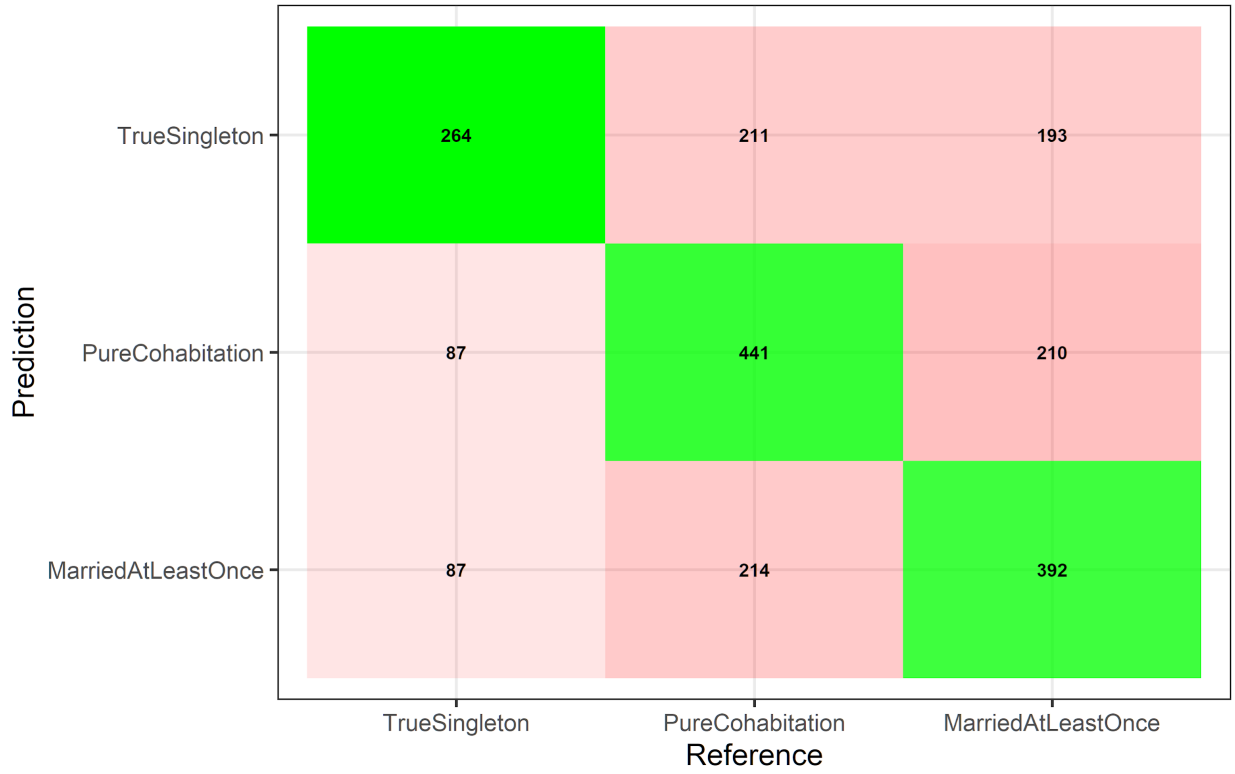
Note. The red line indicates the macro average of the other three curves and is identical to the gbm curve presented earlier in Figure 3.

Figure 5 below shows a confusion matrix for the gbm prediction results using the 2,099 test rows, which as a reminder consisted of 795 Married at Least Once, 866 Pure Cohabitation, and 438 True Singleton cases. As noted in the figure, green shading indicates correct responses and red shading indicates incorrect responses. Darker shading is associated with a greater proportion of responses in a particular bucket with respect to the reference class, with greener shading

indicating greater accuracy and redder shading indicating more misclassifications.

Figure 5

A confusion matrix on the test data for the gbm model in Analysis I.



Green/red indicate correct/incorrect. Darker shading indicates a larger proportion of cases relative to the reference.

Similar to Figure 4, the gbm model showed the strongest classification performance for True Singletons. Performance for True Singletons and individuals who Married At Least Once was similar and slightly worse by comparison. Numerically, balanced accuracy for each of the three classes was .680 (TrueSingleton), .634 (PureCohabitation), and .631 (MarriedAtLeastOnce), with a macro average of .648 as shown earlier in Figure 2.

Variables of Importance

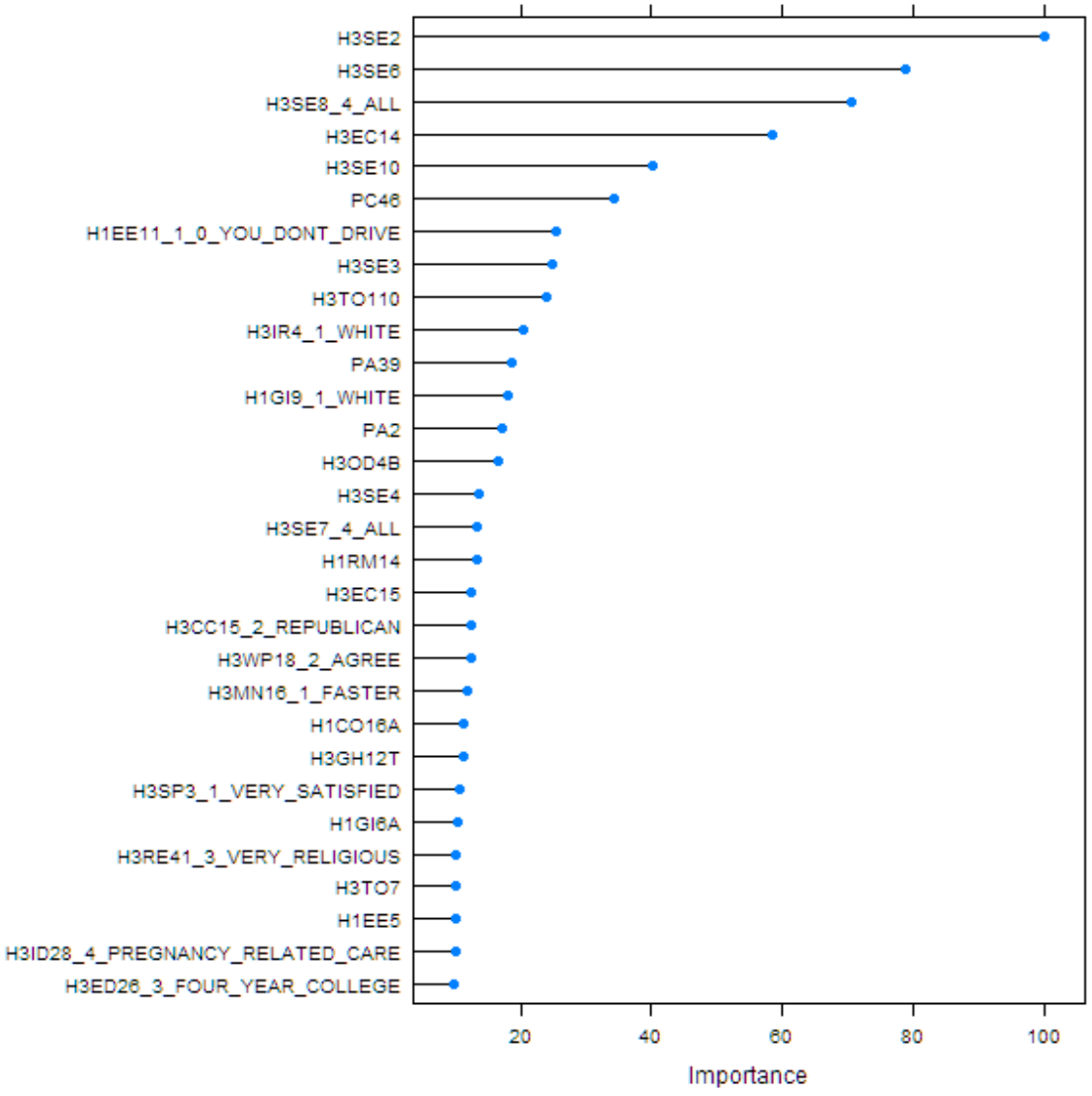
Variables of importance were extracted from the best gbm model to help understand which IVs were most useful for making accurate predictions. Figure 6 below shows the top 30 overall

variables of importance sorted by importance.

Figure 6

The top 30 variables of importance for the best model (gbm) in Analysis I.

To provide additional details on the variables of importance, Table 4 below provides basic details for each variable, including the item text, relevant response (when applicable), the wave it appeared in, the code book it is drawn from, and the corresponding Add Health topic(s). In addition, the table includes information about the key observed pattern for each variable, along with summary statistics for variable importance and the p-value for each item based on a follow-up regression analysis described in the next section. As shown in the table, most of the



top 30 important variables came from Wave III (21), along with nine from Wave I and none from Wave II.

Table 4

Summary details of the top 30 variables of importance for the best model in Analysis I

Item	Item Text	Response	Wave	Code Book	Topics	Key Pattern	Importance	p	Sig.
H3SE2	How old were you the first time you had vaginal intercourse?		3	Sexual Experience and Sexually Transmitted Diseases (STDs)	Sexual Behavior	With greater values: more TrueSingleton and MarriedAtLeastOnce and less PureCohabitation	100.00	<.001	***
H3SE6	How many times have you had vaginal intercourse in the past 12 months?		3	Sexual Experience and Sexually Transmitted Diseases (STDs)	Sexual Behavior	With greater values: more PureCohabitation and MarriedAtLeastOnce and less TrueSingleton	78.82	<.001	***
H3SE8	On how many of these occasions did {YOU/YOUR PARTNER} use a condom?	4 - all	3	Sexual Experience and Sexually Transmitted Diseases (STDs)	Contraception; Sexual Behavior	When selecting this response: more TrueSingleton and less PureCohabitation and MarriedAtLeastOnce	70.57	<.001	***
H3EC14	Do you have an email account?		3	Economics and Personal Future	Computer and Email Access	When selecting this response: more TrueSingleton and MarriedAtLeastOnce and less PureCohabitation	58.35	<.001	***
H3SE10	The most recent time you had vaginal intercourse did {YOU/YOUR PARTNER} use a condom?		3	Sexual Experience and Sexually Transmitted Diseases (STDs)	Contraception; Sexual Behavior	When selecting this response: more TrueSingleton and less MarriedAtLeastOnce	40.16	<.001	***
PC46	Do you think that (he/she) has ever kissed and necked?		1	Parent In-Home Index: Child Specific	Sexual Behavior	When selecting this response: more PureCohabitation and MarriedAtLeastOnce and less TrueSingleton	34.18	<.001	***
H1EE11	About how many miles do you drive each week?	1 - 0, you don't drive	1	Expectations, Employment, Income	Transportation	When selecting this response: more PureCohabitation and less MarriedAtLeastOnce	25.31	<.001	***
H3SE3	With how many partners have you ever had vaginal intercourse, even if only once?		3	Sexual Experience and Sexually Transmitted Diseases (STDs)	Sexual Behavior	With greater values: more PureCohabitation and less TrueSingleton	24.82	<.001	***
H3TO110	During the past 30 days, how many times have you used marijuana?		3	Tobacco, Alcohol, Drugs, Self Image	Illicit Drug Use; Marijuana; Substance Use/Abuse	With greater values: more TrueSingleton and less MarriedAtLeastOnce	24.05	0.07	.

Item	Item Text	Response	Wave	Code Book	Topics	Key Pattern	Importance	p	Sig.
H3IR4	Indicate the race of the respondent from your own observation (not from what the respondent said).	1 - White	3	Interviewer's Report	Race/Ethnicity	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation and TrueSingleton	20.25	<.001	***
PA39	How old were you when you were first married?		1	Parent In-Home Index: Core	Marriage; Parental Background Information	With greater values: more TrueSingleton and less MarriedAtLeastOnce	18.64	<.001	***
H1GI9	Interviewer: Please code the race of the respondent from your observation alone.	1 - White	1	General Introductory	Race/Ethnicity	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation and TrueSingleton	18.05	<.001	***
PA2	How old are you?		1	Parent In-Home Index: Core	Age; Parental Background Information	With greater values: more TrueSingleton and less PureCohabitation	17.18	<.001	***

H3OD4B	What is your race (check all that apply): black or African American		3	Overview and Demographics	Race/Ethnicity	When selecting this response: more PureCohabitation and less MarriedAtLeastOnce	16.60	<.001	***
H3SE4	With how many different partners have you had vaginal intercourse in the past 12 months?		3	Sexual Experience and Sexually Transmitted Diseases (STDs)	Sexual Behavior	With greater values: more PureCohabitation and less TrueSingleton and MarriedAtLeastOnce	13.49	<.01	**
H3SE7	On how many of these occasions of vaginal intercourse in the past 12 months did you or your partner use some form of birth control or pregnancy protection?	4 - all	3	Sexual Experience and Sexually Transmitted Diseases (STDs)	Contraception; Sexual Behavior	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation	13.43	0.01	*
H1RM14	Has she [resident mother] ever smoked cigarettes?		1	Resident Mother	Parental Background Information; Tobacco	When selecting this response: more PureCohabitation and less TrueSingleton	13.28	<.001	***
H3EC15	Do you have a checking account?		3	Economics and Personal Future	Finances/SES	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation	12.45	<.001	***
H3CC15	With which [political] party do you identify?	2 - Republican	3	Civic Participation and Citizenship	Civil/Political Activities	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation	12.39	<.001	***
H3WP18	How much do you agree or disagree with the next statement? You enjoy doing things with your [current residential] mother.	2 - agree	3	Parental Support and Relationships	Parental Support & Relationship	When selecting this response: slightly more PureCohabitation and slightly less MarriedAtLeastOnce	12.29	0.84	
H3MN16	In terms of taking on adult responsibilities, would you say you grew up faster, slower, or at about the same rate?	1 - faster	3	Mentoring	Development	When selecting this response: more MarriedAtLeastOnce and less PureSingleton	11.83	<.001	***
H1CO16A	Have you ever been told by a doctor or a nurse that you had Chlamydia?		1	Contraception—Audio CASI	Illness/Disease; Sexually Transmitted Disease	When selecting this response: more PureCohabitation and slightly less TrueSingleton and MarriedAtLeastOnce	11.30	0.88	
H3GH12T	On days when you go to work, school, or similar activities, what time do you usually go to sleep the night (or day) before?		3	General Health and Diet	Occupation; School Performance/Behavior; Sleep	When selecting PM rather than AM: more MarriedAtLeastOnce and slightly less TrueSingleton and PureCohabitation	11.28	<.001	***

Item	Item Text	Response	Wave	Code Book	Topics	Key Pattern	Importance	p	Sig.
H3SP3	How satisfied are you with your life as a whole?	1 - very satisfied	3	Social Psychology and Mental Health	Development	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation	10.79	<.001	***
H1GI6A	What is your race (check all that apply): white		1	General Introductory	Race/Ethnicity	When selecting this response: more MarriedAtLeastOnce and less TrueSingleton and PureCohabitation	10.49	<.001	***

H3RE41	To what extent are you a religious person?	3 - very religious	3	Religion and Spirituality	Religion & Spirituality	When selecting this response: more TrueSingleton and MarriedAtLeastOnce and less PureCohabitation	10.21	<.001	***
H3TO7	During the past 30 days, on how many days did you smoke cigarettes?		3	Tobacco, Alcohol, Drugs, Self Image	Substance Use/Abuse; Tobacco	With greater values: more MarriedAtLeastOnce and less TrueSingleton	10.01	0.04	*
H1EE5	How much money do you earn in a typical non-summer week from all your jobs combined?		1	Expectations, Employment, Income	Employment Status; Finances/SES	With greater values: more MarriedAtLeastOnce and less TrueSingleton	9.98	0.03	*
H3ID28	What was the main reason for your most recent emergency room visit?	4 - pregnancy-related care	3	Illnesses, Medications, and Physical Disabilities	Childbearing/Pregnancy; Illness/Disease; Injury; Receipt of Health Services; Substance Use/Abuse	When selecting this response: more PureCohabitation and less TrueSingleton and PureCohabitation	9.91	0.03	*
H3ED26	Is this a high school, a two-year college, a four-year college, or a graduate school?	3 - four year college	3	Education	Education Status	When selecting this response: more TrueSingleton and less PureCohabitation	9.83	<.01	**

Follow-Up Analysis

To understand the directionality of each variable of importance, a series of follow-up multinomial log-linear regression models were run using the nnet package in R (Venables & Ripley, 2002). For these analyses, a simple one-way model was set up to determine if a given variable in isolation predicted the dependent variable of lifestyle choice using the actual test data outcomes (i.e., observed rather than predicted responses). The two rightmost columns in Table 4 above show the p-values for each of these follow-up tests, and unsurprisingly the majority were significant given the large size of the dataset. As is apparent from examining the table, many of the variables identified as important by the model were related to one another (e.g., several items were related to sexuality, race/ethnicity, etc.) and were associated with similar, overarching patterns. Given this overlap, rather than discussing each individual effect plot, the following section provides a comprehensive discussion of the results of Analysis I and includes illustrative examples of these effect plots that highlight key patterns. Individual effect plots for every variable in Table 4 are nevertheless provided in Appendix C, and they are ordered relative to the importance shown in the summary table.

Results Discussion for Analysis I

Using a series of machine learning models, this researcher predicted lifestyle choice in Wave IV from a wide variety of predictors in the Add Health dataset across Waves I through III. Overall, a gradient boosted model was the most accurate, correctly predicting test cases over 75% of the time. As shown in the variables of importance in Table 3, this model incorporated a variety of interesting independent variables that warrant further discussion. Because many of the variables were similar to one another, it is possible to group them into sets. To achieve this, the code book from which each item was drawn along with the topic level groupings provided in the Add Health Codebook Explorer (2017) was taken into account. The resulting items category sets are further discussed below and tied into existing literature on marriage, cohabitation, and singlehood. More specifically, this section elaborates on the categories sequentially based on the overall importance of the top item in a given set relative to the variables of importance shown in Table 4 above. Table 5 below provides a summary of the item sets structured in the order in which they are discussed below.

Table 5
Analysis I: Categorization of the top variables of importance from Table 4

Category	Unique Items	Items
Sexual Experiences, STDs, and Health	9	**H3SE2**, H3SE6, H3SE8, H3SE10, H3SE3, **H3SE4**, H3SE7, H1CO16A, H3ID28
Economics and Employment	4	H3EC14, **H1EE11**, H3EC15, H1EE5
Parents	5	**PC46**, PA39, PA2, H1RM14, H3WP18
Tobacco and Marijuana	2	**H3TO110**, H3TO7
Race/Ethnicity	4	H3IR4, H1GI9, **H3OD4B**, H1GI6A

Politics	1	**H3CC15**
Life	3	**H3MN16**, H3GH12T, H3SP3
Religion and Spirituality	1	**H3RE41**
Education	1	**H3ED26**

Note: The associated plots for the items in **bold** are included in the sections below. All other plots may be found in Appendix C. Categories in *italics* are also included in Analysis II below.

Sexual Experiences, STDs, and Health

The first topic grouping - Sexual Behavior, including risky sexual behavior, STDs, and contraception use - contains nine distinct items. They are listed here in descending magnitude: H3SE2, H3SE6, H3SE8, H3SE10, H3SE3, H3SE4, H3SE7, H1CO16A, and H3ID28. The actual text of the items and their effects are summarized in Table 4 above.

Sexual behavior showcased essential links with all three lifestyle outcomes. In general, individuals who delayed or did not have sex were more likely to be singletons. By contrast, more sexual partners increased the likelihood of cohabitation, whereas a conservative number was predictive of marriage. These items suggest that sexual health and behavior as early as an individual's first experience with sexual intercourse are linked to present-day sexual habits and have significant predictive weight in determining lifestyle choice. Unfortunately, recent American research discussing how premarital sex affects marriage timing is sparse. However, a 2003 study by Teachman found that premarital sex reduced marriage rates for women, but not for men, suggesting potential discrimination within sexual history exists and how it has affected women by creating a barrier to entry into marriage.

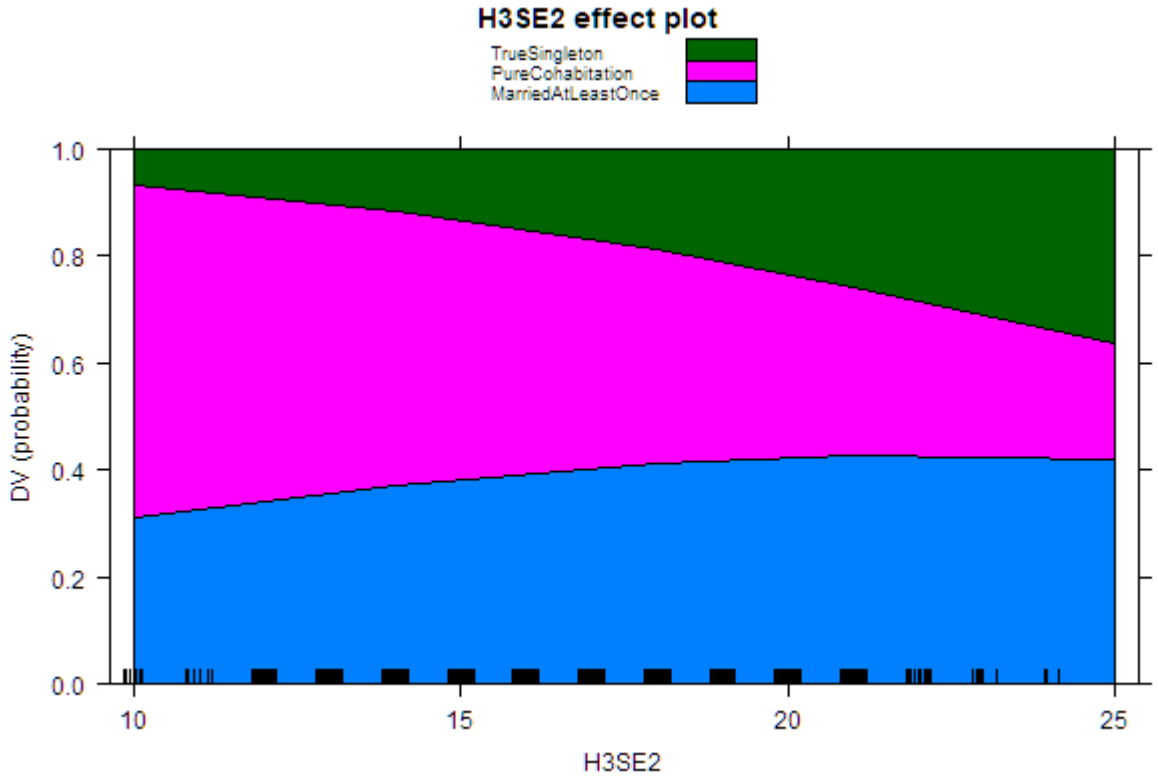
Tying these findings into the existing literature shows that patterns of early sexual behavior appear to have a strong correlation to adult sexual behavior patterns (Halpern et al., 2006). Consistent with this model's findings, increased sexual activity levels before marriage are negatively correlated with marriage and may result in an increased likelihood of cohabitation or

singlehood (Busby et al., 2010). Other research has shown that higher levels of sexual behavior were associated with heightened preference for cohabitation (Willoughby & Carroll, 2010). Regarding singletons, some experimental research has found that an early lack of interest in sexual behaviors is predictive of singlehood, perhaps as a result of personality traits, such as increased autonomy or atypical life goals that place less importance on partnership (DePaulo & Morris, 2005b).

Each of these predicted outcomes shares meaningful relationships with sexual behaviors and attitudes. From the brief number of items gathered in the variables of importance, it appears that sexuality can be examined in a number of important ways to leverage predictions of lifestyle choice. Within the Add Health dataset, if we examine item H3SE6, we find an interesting trend: At the lowest sexual frequency levels, 20% of individuals are true singletons, approximately 25% of respondents are cohabitating, and the remaining percentage of individuals are married. As the frequency of sexual activity over the last twelve months increases, we see the number of true singletons diminish, the number of cohabitation steadily decreases, and the number of married individuals increases until they make up more than 70% of the total respondents.

When we examine the item H3SE2 (see Figure 7 below) and its associated effect plot, we find a relationship between age at first sexual experience and lifestyle choice outcome. The trend exhibited in the plot suggests that having sex very early in life makes marriage less likely, and having sex later in life makes singlehood more likely, while marriage probabilities remain relatively constant for anyone who has sex after the age of 20.

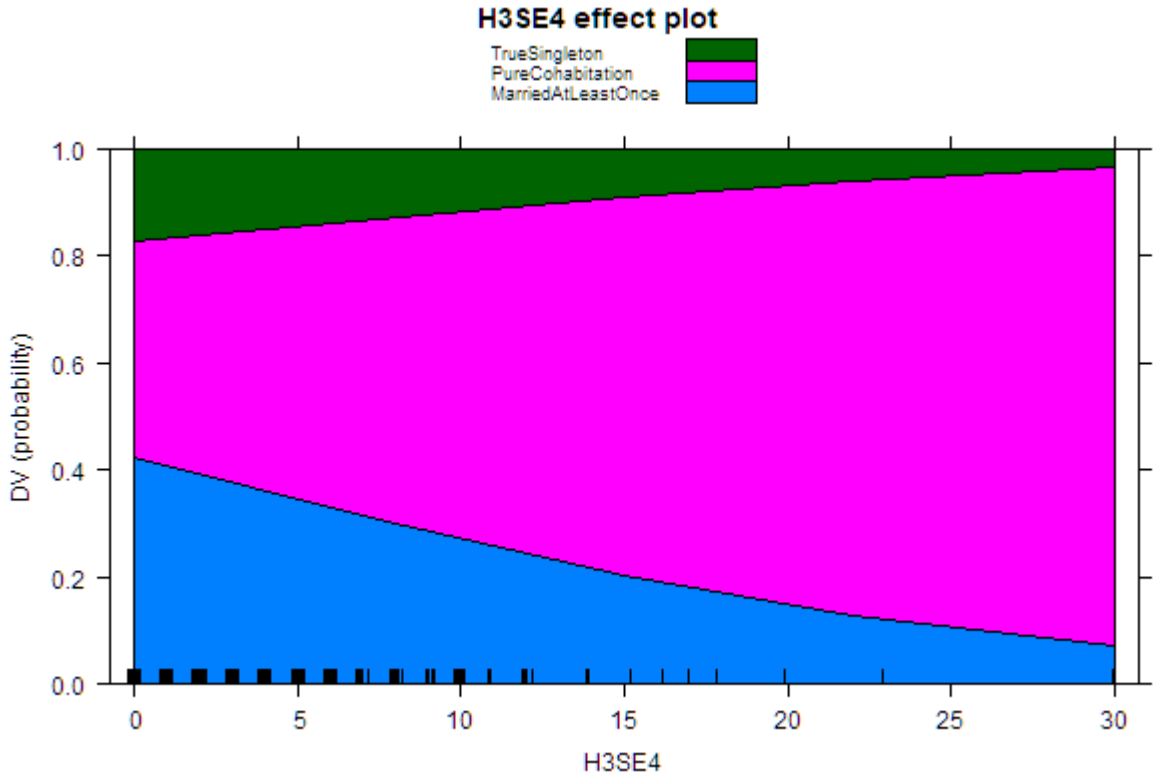
Figure 7
H3SE2



Note. This item asks, "How old were you the first time you had vaginal intercourse?" Age at first vaginal intercourse indicates a high probability of marriage. After the age of 20, the probability for marriage and singlehood increases, while probability for cohabitation decreases.

When we look at H3SE4 (see Figure 8 below), the probabilities skew heavily as the number of partners increases. As the number of partners increases, the probability of cohabitation increases. While partners less than five are nearly equally likely for cohabitation and marriage, partner totals over 20 are highly predictive of cohabitors.

Figure 8
H3SE4



Note. The Wave III item, "With how many different partners have you had vaginal intercourse in the past 12 months?," displays the strong likelihood of cohabitation over marriage and singlehood based upon the number of sexual partners within the past year. The greater the number of partners, the greater the probability of cohabitation.

Economics and Employment

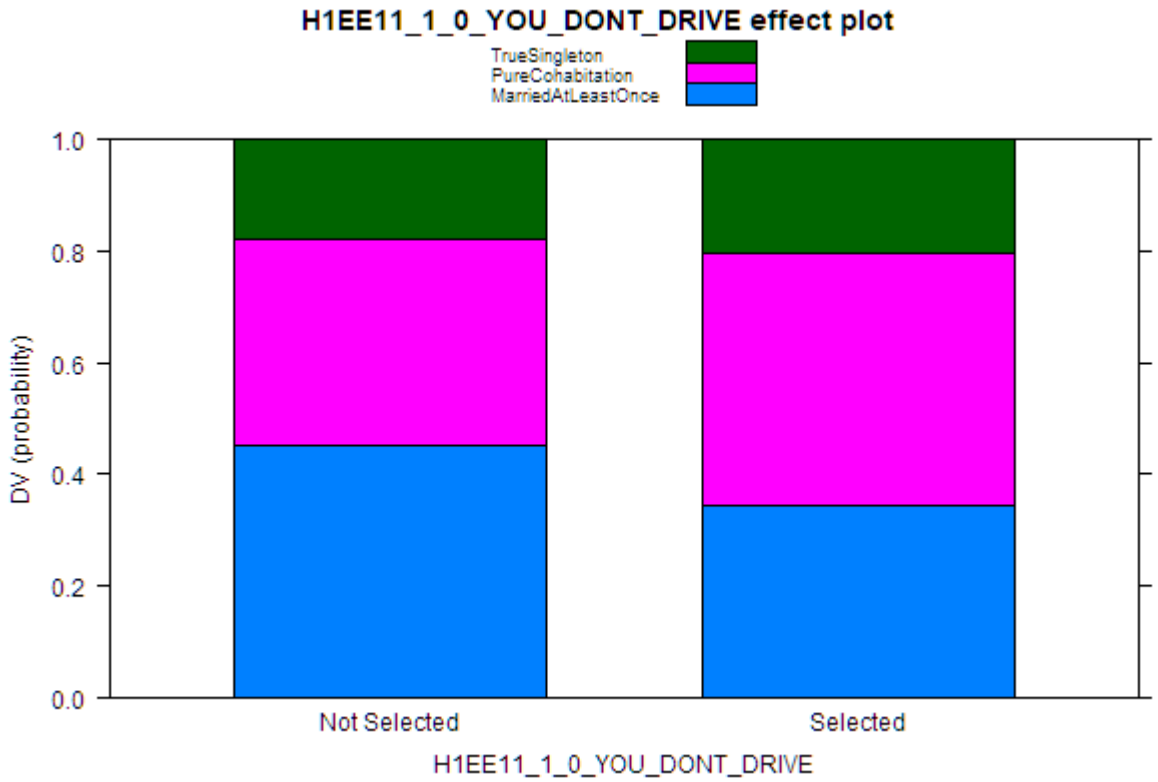
The following items were part of Wave I and Wave II and included information about work and employment-related behaviors and yearly income: H3EC14, H1EE11 (see Figure 9 below), and H3EC15. Two of these items, "Do you have an email account" (H3EC14) and "Do you have a checking account" (H3EC15), point to two of the major economic stratifications in our country: access to the internet and basic banking. If individuals lack access to these two basic services, they will likely experience a variety of other poverty-related stressors. The final question in this set asks about combined wages for non-seasonal work, which can be associated with socioeconomic status (SES). This was demonstrated by Shafer and James (2013), who analyzed data from the National Longitudinal Study of Youth (n = 12,231) in an effort to determine the relationships between SES and marriage timing, which revealed some complex

relationships. Their analysis also concluded that lifelong wealth stratification strongly predicted union formation in various ways, proposing that strongly distinctive economic variables like those described above separate individuals into economic classes and are economic barriers that further influence lifestyle choice.

When we examine the effects plot for item H3EC15, we find that individuals who lack a bank account are approximately 20% likely to be true singletons, 50% cohabitors, and 30% married. Contrastingly, those that had access to a bank account were around 22% likely to be true singletons, 28% cohabitors, and the remaining 50 % are married, further suggesting that access to financial services is strongly related to marriage.

The fact that these items have been identified as important is not surprising since a reasonably standard finding in marriage research is that personal wealth, such as homeownership, is a strong predictor of marriage (Schneider, 2011). The items within this category may also represent preferences in partner selection towards financially stable individuals. In a recent study, Arundel & Ronald (2020) found that, despite homeownership rates approaching an historic low for younger Americans, individuals have the desires of stability and economic freedom that come with owning their home and paying it off rather than renting. With homeownership as a goal for many individuals, recognizing this relationship may help to explain why those who have a mortgage are more likely to get married. Interestingly, research by Greinstein et al. (2014) found that homeownership by single individuals negatively predicted later marriage, but homeownership by couples who were cohabitating predicted increased likelihood of marriage. This compilation of research signifies that homeownership has strong relationships with marriage, cohabitation, and singlehood.

Figure 9
H1EE11



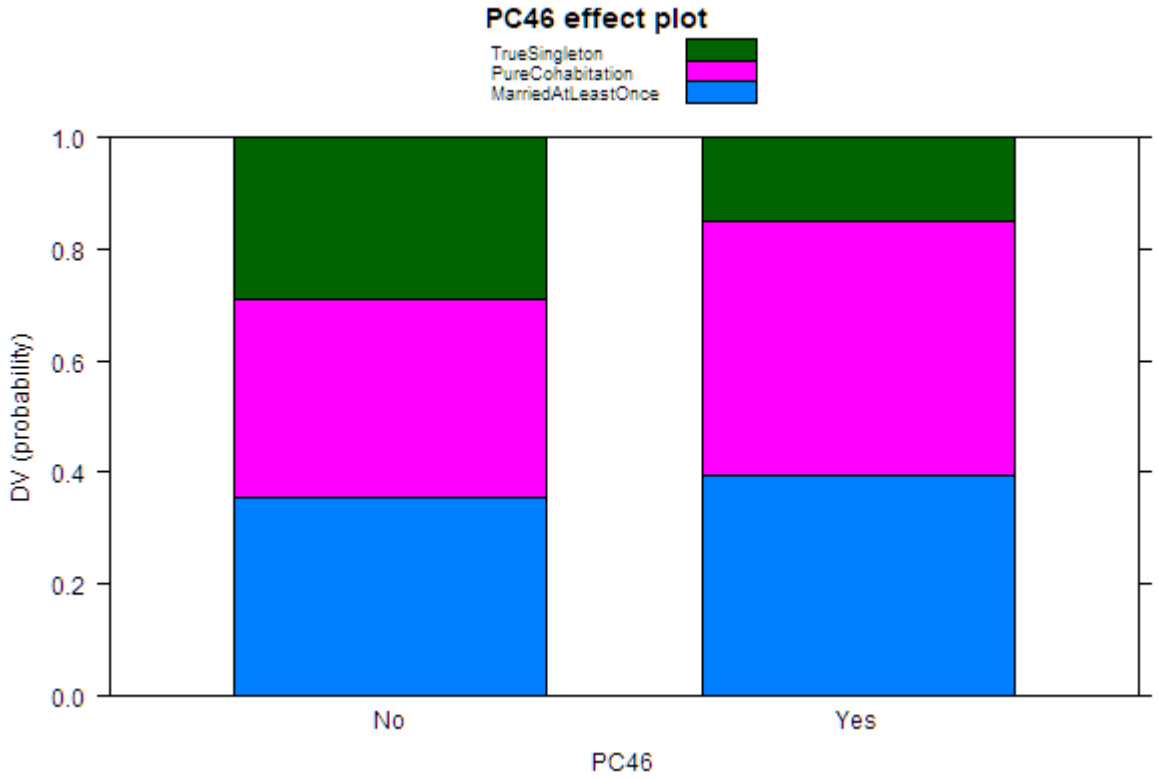
Note. This item asks, "About how many miles do you drive each week?" Selected indicates the respondent does not drive. Individuals who do not drive are more likely to be cohabitors. While Selected also is more indicative of singlehood than Not Selected, there is a minimal difference.

Singlehood may also be affected by low levels of personal wealth and poverty. In many countries, especially in areas where wealth concentration and inequality are high, poverty may be one of the most significant predictors of singlehood (Yamikuski, 2007). Lifestyle choice appears to be affected by personal wealth prior to union formation. Lifestyle choice also subsequently produces different levels of wealth accumulation. Thus, it would seem that lifestyle choice prediction and its inner workings with economics are a topic of potential importance.

Parents

The following items were part of a survey that included asking parents about their children and children about their parents: PC46 (see Figure 10 below), PA39, PA2, H1RM14, H3WP18. These items included questions about attitudes and behaviors, as well as concrete information including information like age when the parents were first married.

Figure 10
PC46



Note. This item asks, "Do you think that (he/she) has ever kissed and necked?" Yes indicates the parent believes their child(ren) has kissed and necked previously and were most likely to be cohabitors, closely followed by marriage.

Figure 10 above shows PC46 and assesses parental monitoring by asking not only if the child is engaging in sexual activity, but also if the parent is aware of it. Di Clemente et al. (2001) conducted a set of studies using logistic regression to examine the relationship between parental monitoring and a variety of outcomes related to sexual activity and risk-taking behavior that could lead to medical harm. Di Clemente et al.'s research found significant relationships between parental monitoring and reduced incidence of sexually transmitted diseases and risky sexual behavior. As previously stated, risky sexual behaviors and unplanned activity around sex early in life had strong relationships with union formation. Parental monitoring early in life may assist in structuring these planful activities that later reduce risky behaviors and strengthen future union

formation.

Item PA39 is important because research has previously suggested that early marriage by mothers is associated with earlier marriage by adult offspring (Thornton, 1991). Thornton's research utilized a life-history analysis method and found that the timing of a mother's marriage had significant effects on the union formation timing for her adult children's cohabitation and marriage. Item H1RM14 is important from the standpoint of parental modeling of health behaviors and a planned behavior standpoint. Harakah et al. (2004) conducted a longitudinal study utilizing structural equation modeling (SEM) to analyze data from 1,070 individuals. Researchers found that choices, such as smoking as a parent, had strong relationships with planning to model healthy behaviors for children, which in turn strongly related to children's health behavior histories.

One particularly striking example was found when examining the effect plot of PA39, an item that asks a parent how old they were when they got married. The younger the parent, the greater the likelihood their offspring became married. In fact, at the youngest recorded parental marriage ages, nearly 50% of all adult children were married, nearly 40% were cohabitors, and around 10% were singletons. As we move along the x-axis of the graph towards increasing age of parental marriage, we find a sharp decline in married adult children, a slightly less step decrement in rates of cohabitation, and a sharp increase in adult children who are true singletons such that in the oldest parental marriages adult children are true singletons 60% of the time. Of all the findings in this study, increased age of parents and first marriage is associated most strongly with an increased incidence of true singletons.

Interestingly, past research on marriage has shown the opposite effect: high levels of parental supporting behaviors, including financial support, lead to increased marriage and

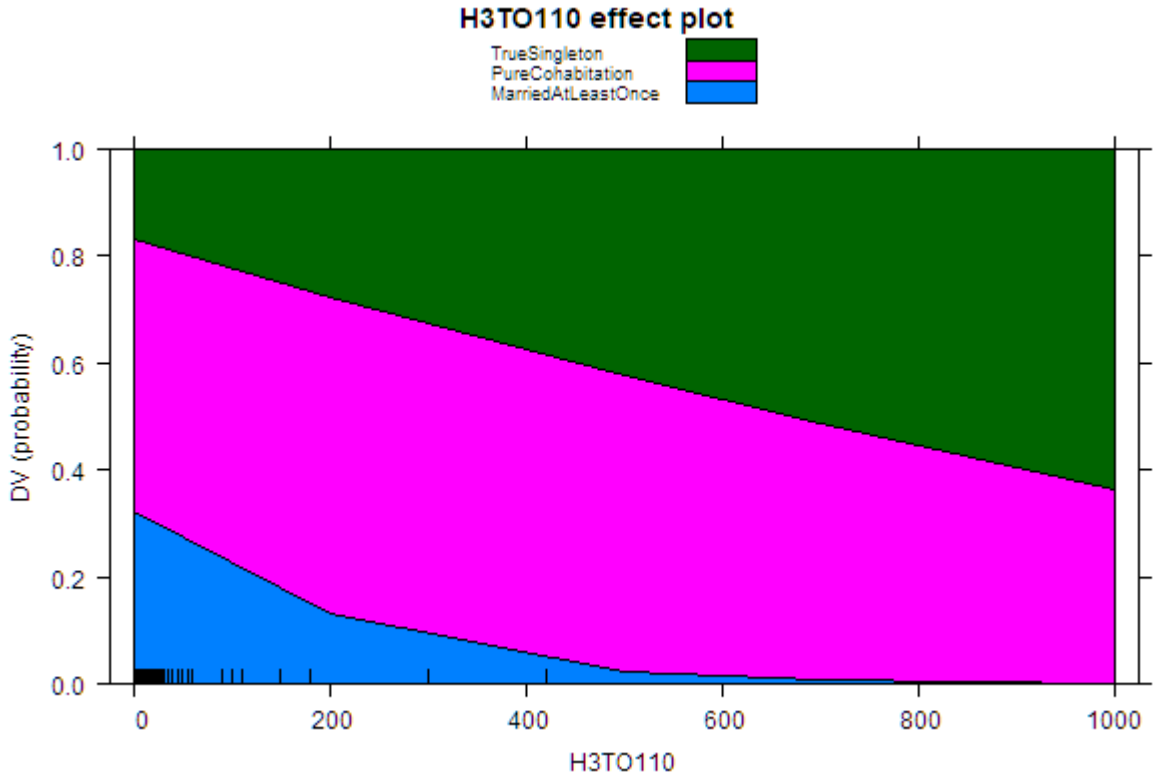
marriage stability (Swartz et al., 2011). Additionally, parental attitudes have substantial effects on adult children's choices to marry or cohabitate (Axinn & Thornton, 1993; Hall, 2006).

Additionally, poor parental support and neglect can lead to isolation in children with long-term biological consequences for the inflammatory system, leading to psychological and behavioral distress that may create barriers to entry into unions with others, either in marriage or cohabitation. Therefore some individuals lacking parental social support may find it difficult to form partnerships (Lacey et al., 2014).

Tobacco and Marijuana

This item set contains two items regarding the frequency of marijuana and cigarette use: H3TO110 and H3TO7. When examining the Add Health data for item H3TO110 (see Figure 11 below) we find that marijuana use in the last 30 days creates important distinctions between lifestyle choice outcomes. As marijuana use increases, the number of married individuals drops nearly to zero and cohabitors represent nearly 40% of the sample with the remainder being true singletons. The data suggests that greater frequency of marijuana use over a thirty day period is negatively associated with marriage and positively associated with cohabitation and true singletons.

Figure 11
H3TO110



Note. This continuous item asks, "During the past 30 days, how many times have you used marijuana?" The probability of singlehood steadily increases as the number of uses increases, while the probability of marriage rapidly drops off. The rug lines along the x-axis show that the majority of individuals used less than approximately 50 times, although there is clearly a long tail in which the general pattern holds.

Some research suggests that early adolescent marijuana use causes significant disruption in late marriage rates (Menasco & Blair, 2004), which is consistent with the model's findings. Additionally, loneliness and the accompanying psychological distress often lead to self-medicating through substance use (Segrin et al., 2018), and thus their formation of significant marriage relationships may be delayed.

Hoffmann (2018) found that there was a strong relationship between marijuana use and cohabitation by using longitudinal methods to analyze data from three waves of the National Survey of Youth and Religion (n = 2,202) from the years 2002 through 2008. Hoffmann was able to identify that, prior to cohabitation, marijuana use predicted cohabitation and among female participants cohabitation was strongly related with increased marijuana use. With regard

to tobacco smoking. Chipporie et al (2001) analyzed data from the Current Population Survery's subsections related to tobacco use. Analysis of the data using assortative economic modeling to explore the maximum matching outcomes while accounting for heterogeneities regarding variables from the tobacco items in the aforementioned survey. This research by Chioppori et al (2001) found that there is a strong preference associated with smoking status when it comes to selection of marital partners and that discordant desires for smoking partners was exacerbated by the fewer number of women who smoked compared to the number of men who smoked. This research suggests that partner preferences and available supply of smokers and non smokers exerts pressure on the marriage market outcomes for couple formation The results suggest that substance use including tobacco and marijuana have important relationships to health behaviors that regulate lifestyle choice when selecting partners.

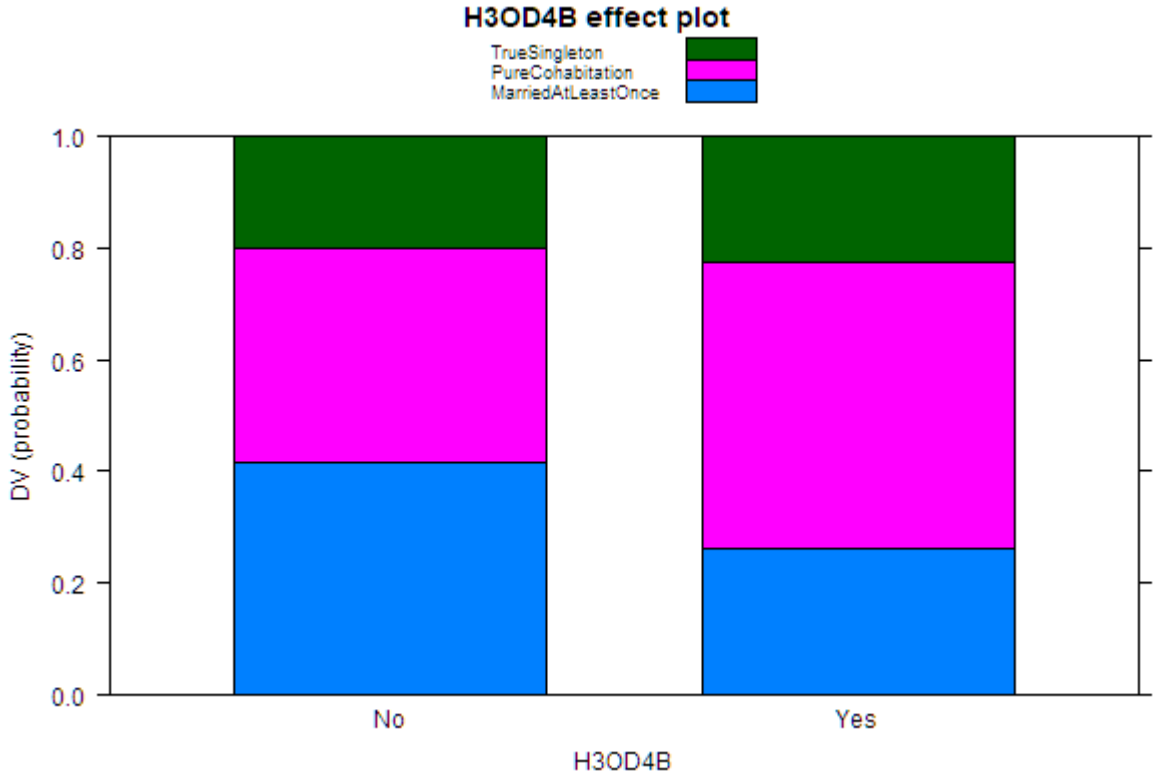
Race/Ethnicity

All four items that make up the itemset covering the topic of race and ethnicity that appear in the variables of importances analysis in Analysis I have to do with being African American or Caucasian. The items include the following: H3IR4, H1GI9, H3OD4B (see Figure 12 below), and H1GI6A.

The findings related to the prediction of lifestyle choice and race/ethnicity are complex to interpret and are likely affected by bias at many levels. For instance, research has found that economic opportunity plays a significant role in explaining the difference between Black and White marriage patterns in American society. Any analysis of the topic must seriously consider the sociological stratifications that each demographic group is subject to (Bennet et al., 1989). More specifically, research suggests that marriage is avoided by low SES African Americans out of necessity because of costly financial barriers (Manning & Smock, 1995). Additionally, the

same research by Manning and Smock found that African Americans frequently cited cohabitation as a transition that delayed marriage costs.

Figure 12
H3OD4B



Note. This item asks, "What is your race (check all that apply)." For this effect plot, Yes indicates the respondent chose Black or African American, while No indicates they chose a different item. Individuals who did not identify as Black or African American were more likely to be married, while Black or African American individuals were more likely to cohabit. The likelihood of singlehood is relatively equal for both Yes and No.

Caucutt et al. (2018) found that 83% of Caucasian women between 25 through 54 years of age were ever married. While at the same time, 56% of age comparable African American women were married. Caucutt's finding was clear, with a significant separation of 27%. Caucutt surveyed a variety of important social, economic, and historical reasons that included differential incarceration rates (higher for African Americans) structural poverty rates that disfavored African Americans. Specifically, these economic factors included reduced homeownership rates, fewer opportunities for education, and reduced employment opportunities. Caucutt ultimately

concluded that marriage is both delayed and reduced for African Americans and stated that marriage is inherently riskier, and accompanied by fewer advantages, for African Americans than Caucasians in the US.

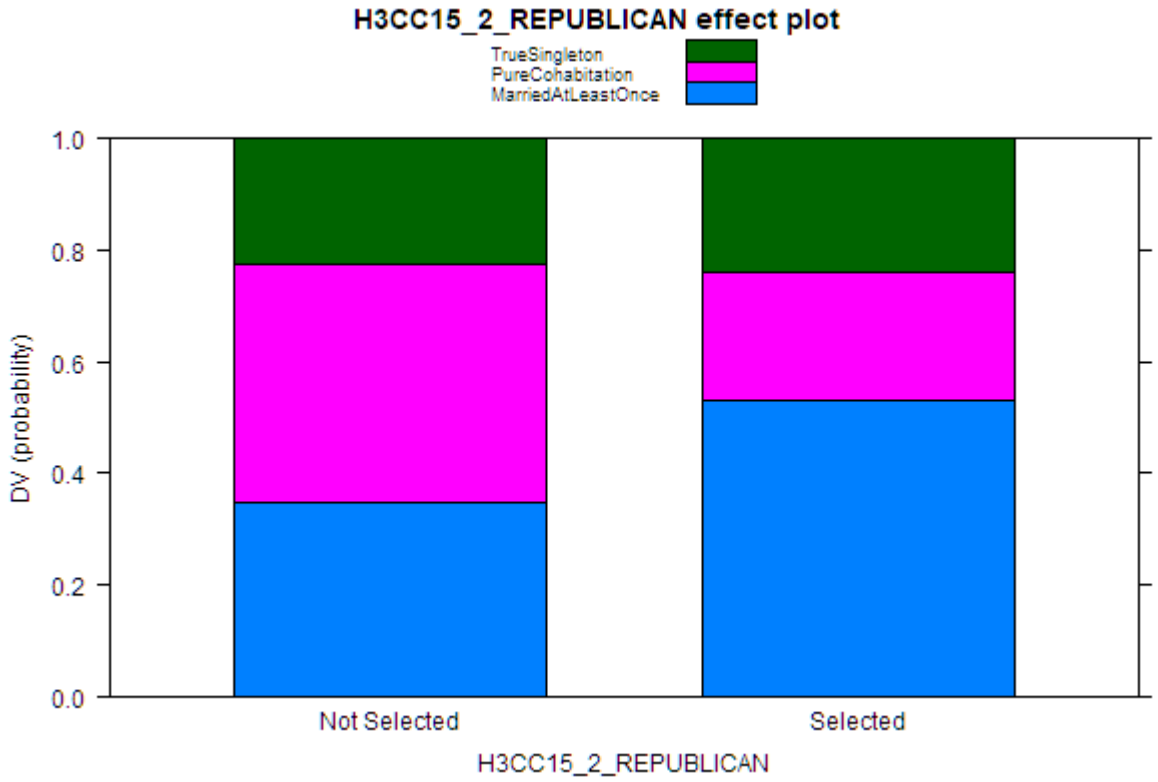
With regard to singlehood, the differential threats and harms experienced by African Americans have significant implications. Some research has examined household disruption; often, the result of separation, legal divorce, work migration, and incarceration experienced by African Americans during childhood can affect attachment style (Underwood, 2013). In turn, these attachment style differences can lead adult African Americans to experience higher levels of relationship dissatisfaction and loneliness. Additionally, greater levels of interpersonal violence and bullying experienced by African Americans often predict adult reports of loneliness in a way that negatively impacts physical and mental health (Segrin et al., 2002). These variables of importance related to perceived and self-identified racial/ethnic identity are strongly predictive of a lifestyle choice that appears to have serious health consequences for African Americans and warrants further, finer-grained research.

Politics

This category includes a single item: H3CC15 (see Figure 13 below), “With which [political] party do you identify?” A 2009 Gallup poll interviewed 29,000 individuals, and the resulting information regarding political affiliations and marital status was analyzed (Newport, 2009). The 29% of individuals who identified as republican varied between married and unmarried individuals: 33% married and 22% unmarried. This was contrasted with rates among the democrats who represented 35% of all respondents: 31% married and 41% unmarried. These findings suggest that marriage rates differ significantly between self-identified republicans and democrats in America. Although these findings are not the result of a predictive experiment, they

represent a demographic trend that can be further explored by using item H3CC15 as a predictor of lifestyle choice.

Figure 13
H3CC15



Note. This item asks, "With which [political] party do you identify?" Selected means the respondent identified as Republican, while Not Selected indicates the respondent selected an alternate choice. Republicans are far more likely to be married, while individuals who did not identify as Republican were more likely to be cohabitors.

When examining the effect plot for the item H3CC15, we find that participants who endorsed republican as their political affiliation were around 20% more likely to be married than non-republican affiliated participants. The number of true singletons was identical across answers to the item.

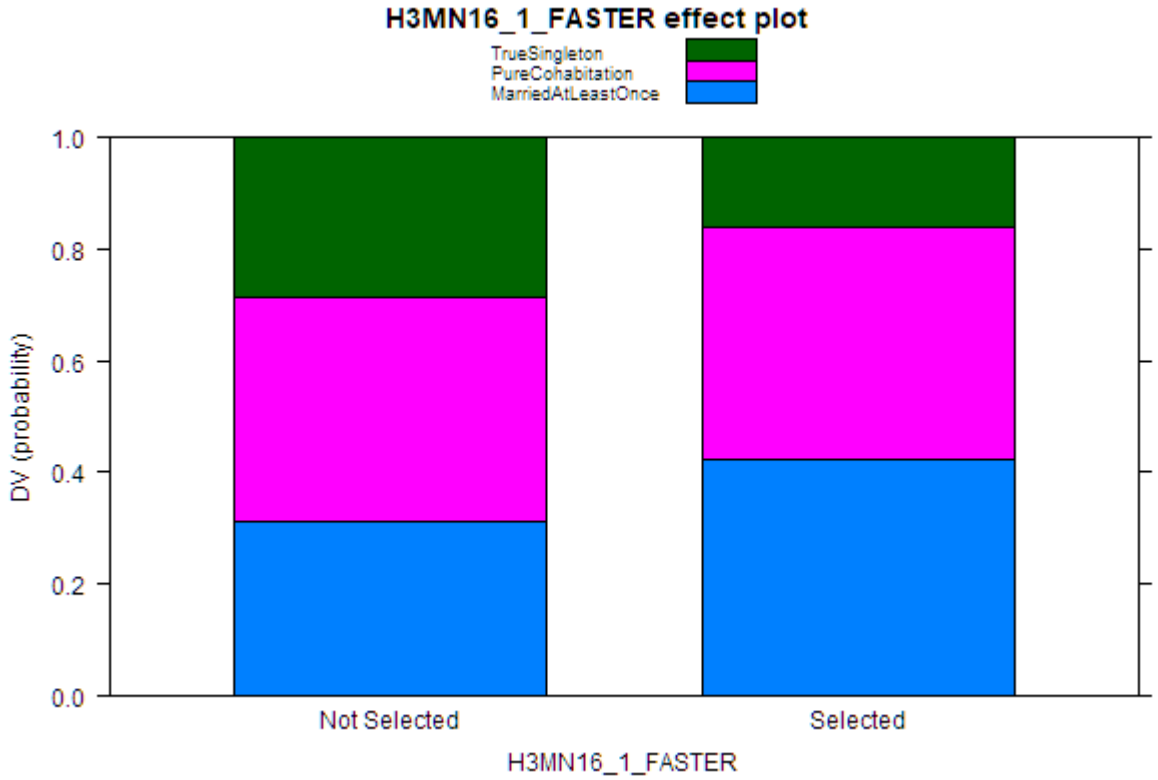
Life

This set contains items from the topic construct Life, which the Add Health researchers intended to explore attitudes and opinions held by participants about their lives in general terms:

H3MN16 (see Figure 14 below), H3GH12T, and H3SP3. One item refers to recollections across childhood about the rate of taking on responsibility, and a second assesses current perceptions of life satisfaction. Finally, another item asks for concrete terms about sleep the night before a scheduled responsibility.

The first item in the set (H3MN16) states, “In terms of taking on adult responsibilities, would you say you grew up faster, slower, or at about the same rate?” This item potentially suggests that individuals who experienced childhoods in which they matured rapidly and took on responsibilities may be more likely to marry. Another interpretation is that these individuals feel a need to step up to fulfill a deficit in their home due to some disruption in their family structure that may result in a delay in marriage or cohabitation, or even sidetrack relationships significantly enough to lead an individual to singlehood.

Figure 14
H3MN16



Note. This item asks, "In terms of taking on adult responsibilities, would you say you grew up faster, slower, or at about the same rate?" Selected indicates the respondent felt they grew up at a faster rate. Individuals who felt they grew up at a faster rate were more likely to be married, while the probability of marriage, cohabitation, and singlehood (with a slight higher chance for cohabitation) was roughly equal for individuals who did select Faster.

Johnson and Mollborn (2009) examined Waves I and III using multivariate analyses to examine how a variety of childhood hardships affects feelings of maturity and how this relationship changes with age. They found that a variety of challenges, like parental divorce, poverty, and illness, led younger individuals to feel a greater "subjective age" than their actual chronological age, but they also found that adulthood moderated most of these feelings of having grown up quickly. Research by Eliason et al. (2015) utilized a hierarchical latent class analysis to examine participants from the Youth Developmental Study. The longitudinal sample examined individuals from the ages of 17 to 30 years of age and allowed Eliason et al. to study the transition from adolescence to various adult roles, including relationship formation. They found

that subjective sense of timing and the sense of taking on responsible roles early led to empirical adult role acquisitions like family formation and career development.

The third item in the set (H3SP3) states, “How satisfied are you with your life as a whole?” This item is interesting because it suggests that subjective well-being can predict lifestyle choice. Research on how life satisfaction predicts lifestyle choice is somewhat sparse; however, research by Stutzer and Fry (2005) attempted to determine if happier individuals are more likely to opt for marriage when compared to individuals who reported being less happy. Stutzer and Fry utilized a longitudinal sample (n = 15,268) individuals surveyed at multiple intervals from the years 1984 to 2000. The researchers were able to determine that greater levels of reported life satisfaction were correlated with a higher likelihood of marriage and earlier marriage. Additionally, a more recent study by De Neve et al. (2013) surveyed the findings on the objective effects of subjective well-being and found that happier individuals were more likely to form romantic relationships and get married compared to individuals who reported lower levels of subjective well-being.

Religion and Spirituality

This category is also a single-item set: H3RE41 (see Figure 15 below). A reasonably comprehensive study from 1992 examined parental religiosity, childhood religiosity, and their effects on attitudes and subsequent entry into cohabitation or marriage, which found that individuals from religious backgrounds married at higher rates (Thornton et al., 1992). Thornton et al. used a panel study of mothers and their children to examine the causal relationships between religious commitment and participation in cohabiting relationships and marriages. Thornton et al.’s work revealed that the religiosity of individuals and their parents had significant effects on lifestyle choice. Explicitly, higher levels of religiosity, both endorsed and practiced

within families, correlated with higher levels of marriage. Lower levels of endorsed and practiced religion led to increased levels of cohabitation. Finally, Thorton et al. examined a number of reciprocal relationships and found that high religiosity marriages increased the religiosity of couples, and cohabitation led to decreases in religiosity. Thornton et al.'s insights indicate two divergent trends where low religiosity leads to cohabitation and decreased religiosity and, on a separate track, religious individuals marry to reinforce their religiosity.

Similarly, in a chapter on religion and entry into cohabitation and marriage, Lehrer (2000) stated that religion is a robust predictor of marriage, especially when two partners share religious identification and a similar level of religious commitment. Additional research on religion and cohabitation specifically found that lower degrees of endorsed religiosity was related to an increased likelihood of cohabitation (Cohan & Kleinbaun, 2002).

When examining the effects plot for item H3RE41 we find differences in lifestyle choice based on endorsement of religiosity. Those who do not endorse religiosity are around 20% likely to be singletons, 42% cohabitors, and 38% married. Whereas the religiously-identified are roughly 35% likely to be singletons, 25% cohabitors, and 40% married. Although the difference between married and unmarried in the two groups differs slightly, there is a major increase in singletons in the religious group and a reduction in the number of cohabitors. This finding is similar to the research described above that religiously identified individuals are more likely to avoid cohabitation.

Figure 15

H3RE41

Note. Item H3RE41asks, "To what extent are you a religious person?" Selected indicates Very Religious. Very Religious individuals were far more likely to be married or single than to be cohabitors. However, individuals who did not identify as Very Religious were more likely to be cohabitors or married, with a lower probability of singlehood.

Education

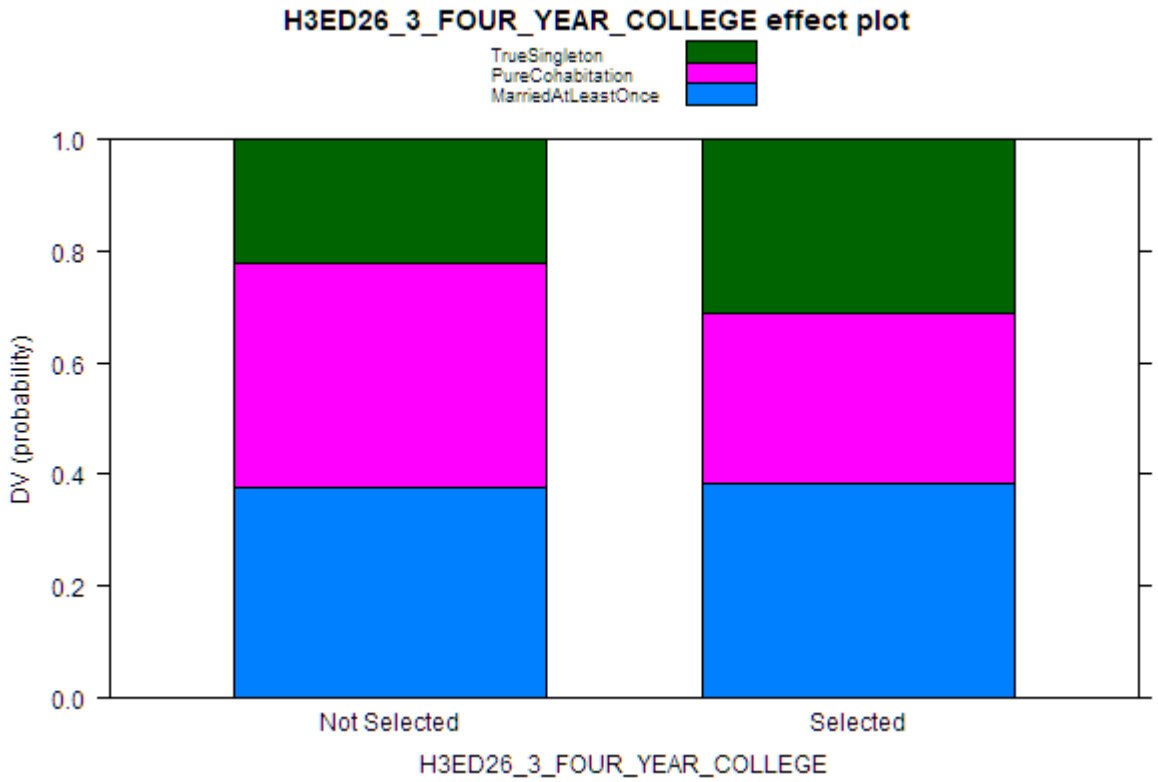
Individuals who indicated they were currently in school during Wave III were presented with a follow-up question (H3ED26): "Is this a high school, a two-year college, a four-year college, or a graduate school?" As shown in Figure 16 below, individuals who were currently attending a four-year college were more likely to be singletons and less likely to be cohabitating at Wave IV. Regardless of their answer to this item, however, the overall marriage probability was high (i.e., ~40% in both cases), suggesting that individuals who were in school at Wave III were generally highly likely to be married at the following wave.

A number of studies have explored the relationship between educational attainment and marriage rates. For example, a study by Musick et al. (2012) examined participant data (n =

3,208) from the National Longitudinal Survey of Youth using a propensity score approach to create stratifications of men and women based on educational attainment. Results indicated that education strongly influenced the timing of first marriage and that these effects were strongest for the minority of the most economically-advantaged individuals. The researchers concluded that education strongly affects marriageability, but is mediated by other inequalities that determine education quality and level of degree (i.e., associates degree, college degree, professional or advanced degree). It would be interesting to see how outcomes on this item change using lifestyle choice outcomes at Wave V.

More recent research by Lichter et al. (2019) examined five years of data from the American Community Survey to determine if a mismatch between women and marriageable partners could account for the decline of marriage. A number of imbalances in female education rates compared to male education rates at a number of geographic levels, including towns, states, and national levels, suggest that a marriage market mismatch exists. In other words, educational attainment by many women outpaces that of men. Therefore, Lichter et al. concluded that the marriage rate has gone down due to an increasing share of unmarried men.

Figure 16
H3ED26



Note. For individuals who were currently in school during Wave III, this item asks, "Is this a high school, a two-year college, a four-year college, or a graduate school?" Individuals who indicated they were currently attending a four year college were more likely to be singletons and less likely to be cohabitating.

Analysis II

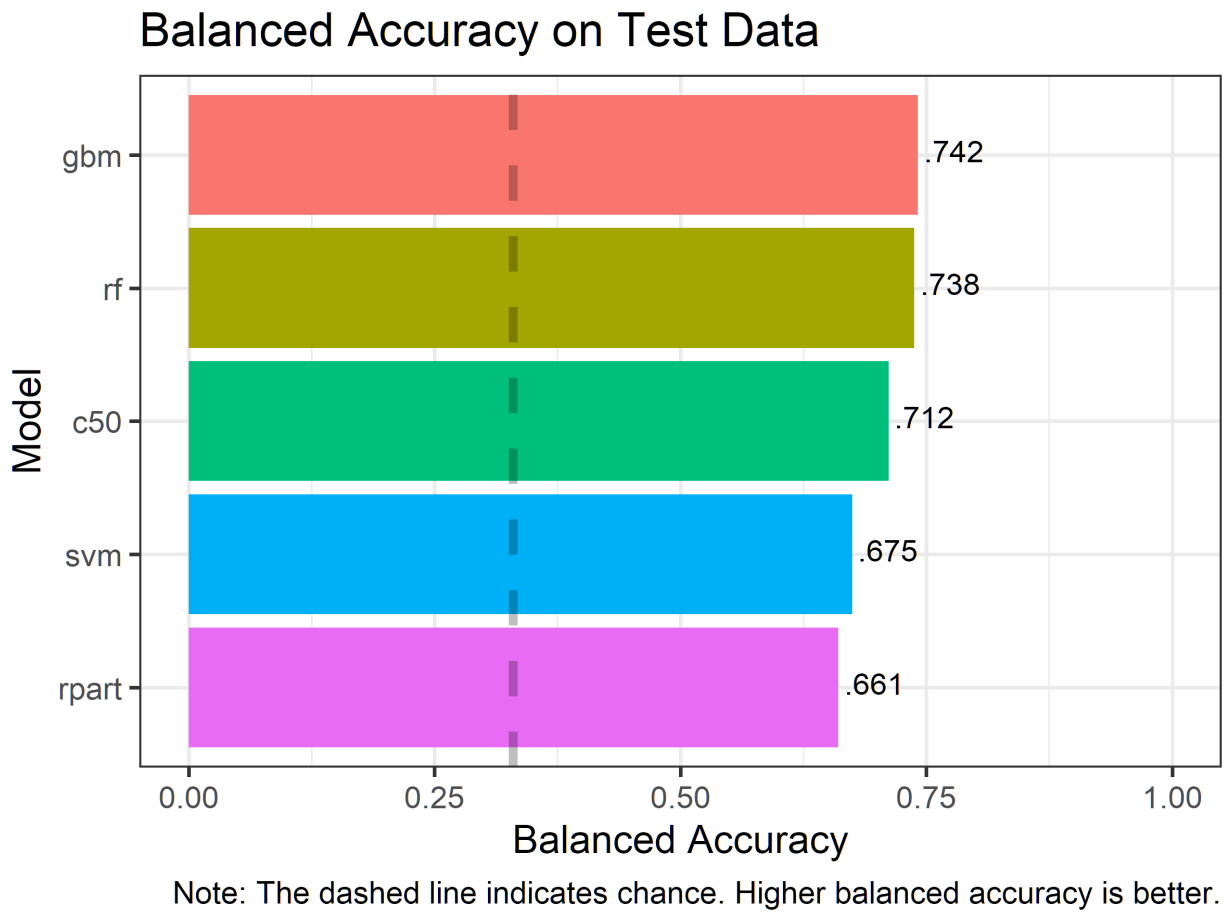
To complement Analysis I, a second analysis was run that focused on descriptive characteristics of lifestyle choice, as discussed in the Defining the Analysis section. In terms of implementation, Analysis II followed a similar approach to Analysis I: for example, both involved testing five types of machine learning models and used identical data cleaning and preprocessing strategies, five-fold cross-validation with grid-search hyperparameter tuning, etc. The key differences in Analysis II are two-fold: 1) predictors from Waves I through Wave IV were included to predict lifestyle choice at Wave IV; and 2) individuals were not filtered from the dataset based on their marriage status prior to Wave IV as in Analysis I.

For Analysis II, the 15,662 rows of the dataset were similarly split into 80% training (12,531 rows) and 20% testing (3,131 rows). After data cleaning and preprocessing, there were 6,237 rows and 9,566 independent variables (IVs) in the downsampled training set, with 2,079 cases each of TrueSingleton, PureCohabitation, and MarriedAtLeastOnce. The 3,131 rows in the 20% testing set contained 1,559 Married At Least Once, 1,052 Pure Cohabitation, and 519 True Singletons (i.e., 3,131 rows in total).

Model Comparison

Consistent with Analysis I, the performance of the five models was evaluated using the test data with macro balanced accuracy on the three-class outcome. The results of this comparison are shown below in Figure 17. Once again, performance for all the models was well-above chance, and unsurprisingly the models performed better in this analysis than in Analysis I - a result attributable to the inclusion of Wave IV data and additional cases that were not filtered based on marriage status at earlier waves. Similar to Analysis I, the best model was the gbm with a balanced accuracy of .742, meaning the model was correct nearly three out of four times on average.

Figure 17
A comparison of balanced accuracy in the training data for Analysis II.

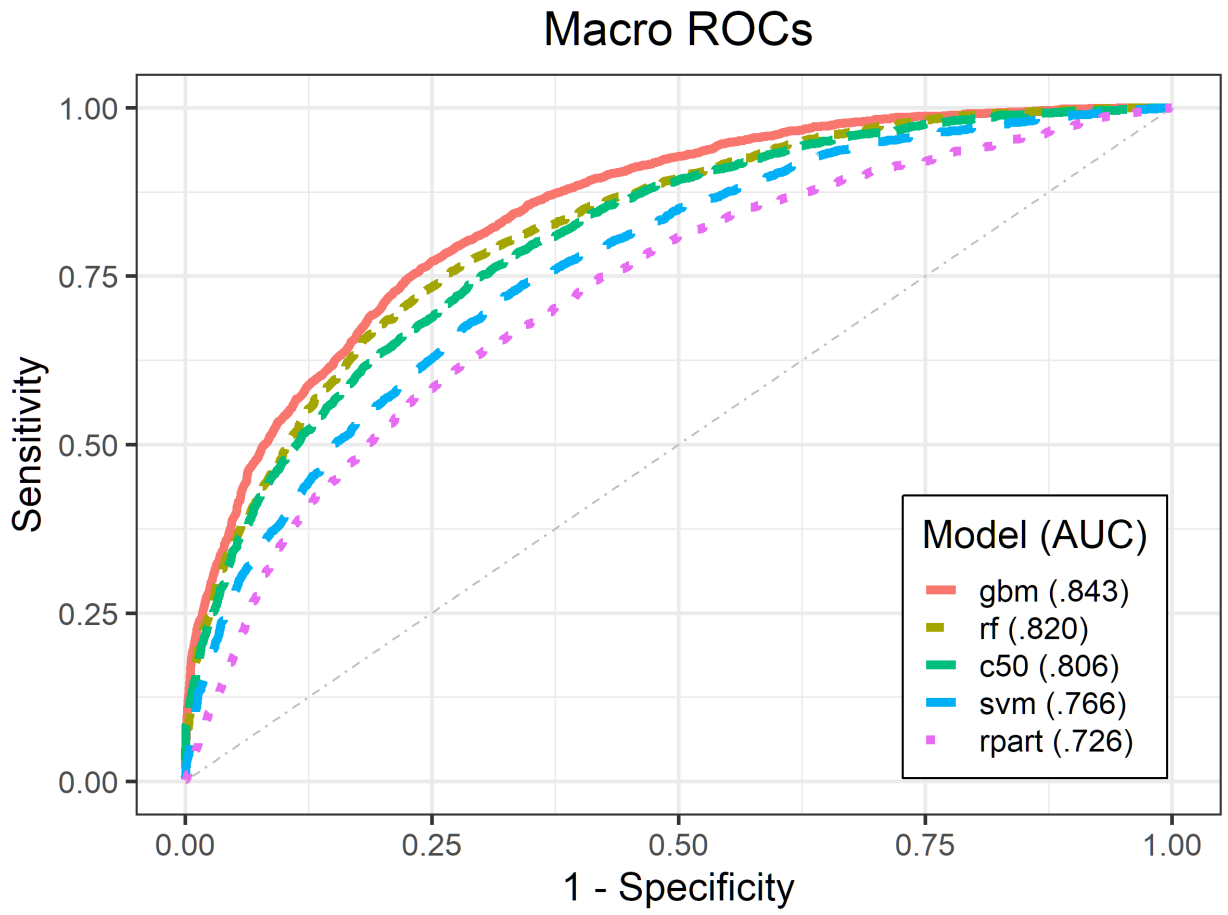


Note. The gradient boosted model (GBM) performed best overall and well above chance. The dashed line indicates chance. Higher balanced accuracy is better.

In addition, macro averaged ROCs and AUCs were calculated for each of the models and plotted for comparison. These results are shown in Figure 18 below and are consistent with Figure 3 above in Analysis I: the gbm model performed best overall in Analysis II and will be used for subsequent analysis and interpretation. Similar to macro balanced accuracy, AUC scores were higher in Analysis II due to the additional data used in this analysis, which provided uplift.

Figure 18
Macro averaged receiver operating curves and AUC scores for each model in the test data for

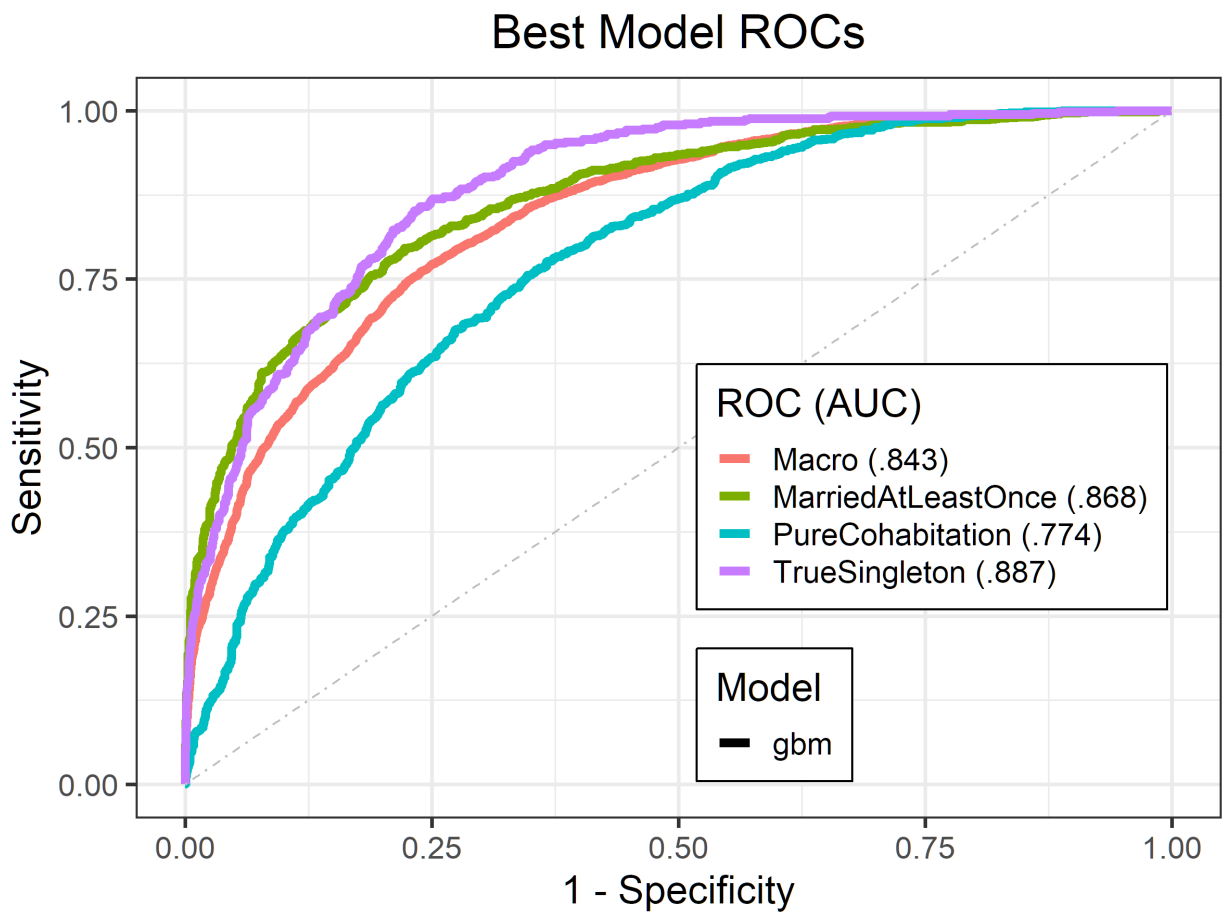
Analysis II.



Investigating the Best Model

Figure 19 below breaks down the ROC curve of the gbm in Figure 18 above (red line) into constituent parts. Once again, the model performed best when predicting True Singleton status (purple line), although this time performance was relatively close for predicting Married at Least Once (green line). By contrast, the ROC for Pure Cohabitation (blue line) exhibited sharply lower performance compared to the other two outcomes that was well below the macro average - a result that was more similar to Analysis I.

Figure 19
The macro ROC curve for the best gbm model in Analysis II and the constituent ROC curves for each class within the dependent variable.

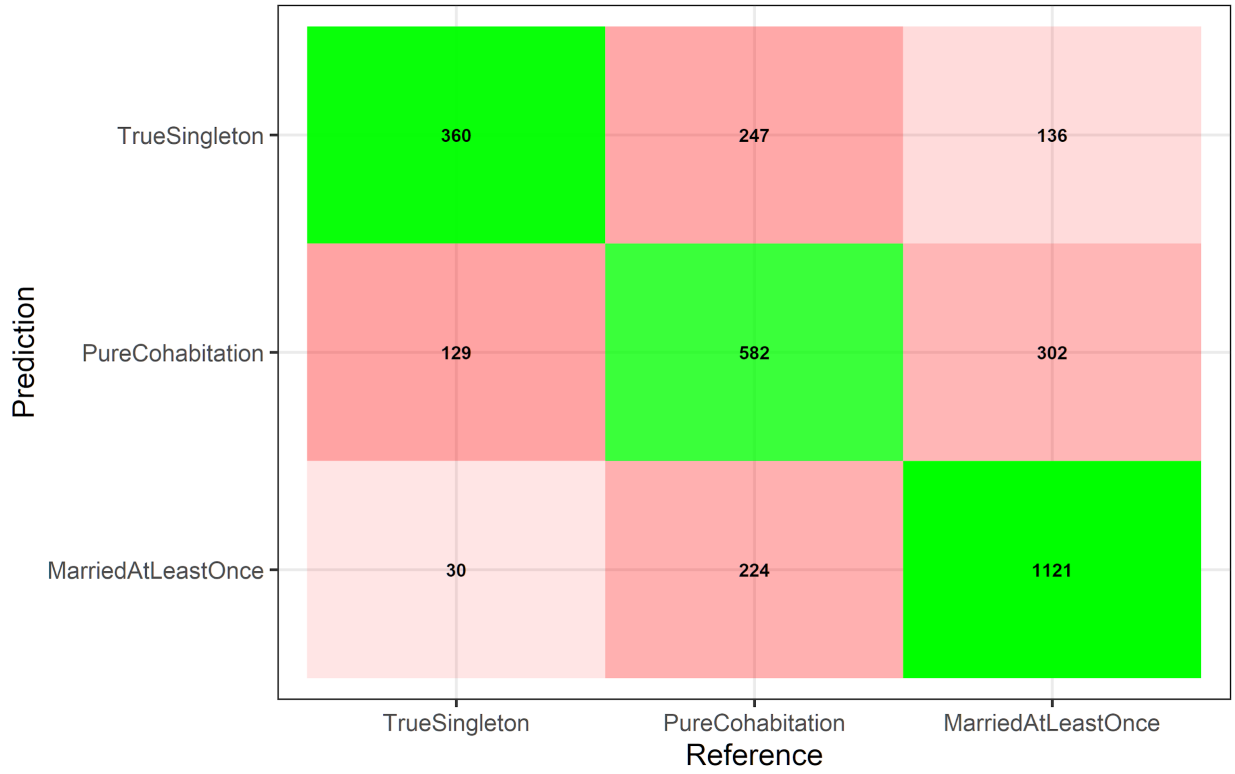


Note. The red line indicates the macro average of the other three curves and is identical to the gbm curve presented earlier in Figure 18.

Figure 20 below shows a confusion matrix for the gbm prediction results using the 3,131 test rows, which consisted of 1,559 Married at Least Once, 1,053 Pure Cohabitation, and 519 Pure Singleton cases as described at the outset. As shown above in Figure 19, the overall balanced accuracy was .742, with balanced accuracy on Married At Least Once (.779) and True Singleton (.774) at the high end and Pure Cohabitation (.673) at the low end. These results are in line with the individual ROCs shown in Figure 19 above.

Figure 20

A confusion matrix on the test data for the gbm model in Analysis II.



Green/red indicate correct/incorrect. Darker shading indicates a larger proportion of cases relative to the reference.

Variables of Importance

Figure 21 below shows the variables of importance for the gbm model in Analysis II visually. These are sorted in descending order of importance, with higher importance near the top.

Similar to Analysis I, Table 4 below provides additional information regarding each variable of importance for subsequent interpretation. This table also shows the p-values from the follow-up multinomial log-linear regression models described previously in Analysis I. The bulk of the variables considered important in this analysis came from Wave IV (25), whereas a small number came from Wave I (3) and Wave III (2). Once again, rather than plotting and describing the effect of each variable one-by-one, the following discussion section provides a summary of

the variables of importance by grouping them in terms of similarity.

Figure 21

The top 30 variables of importance for the best model (gbm) in Analysis II.

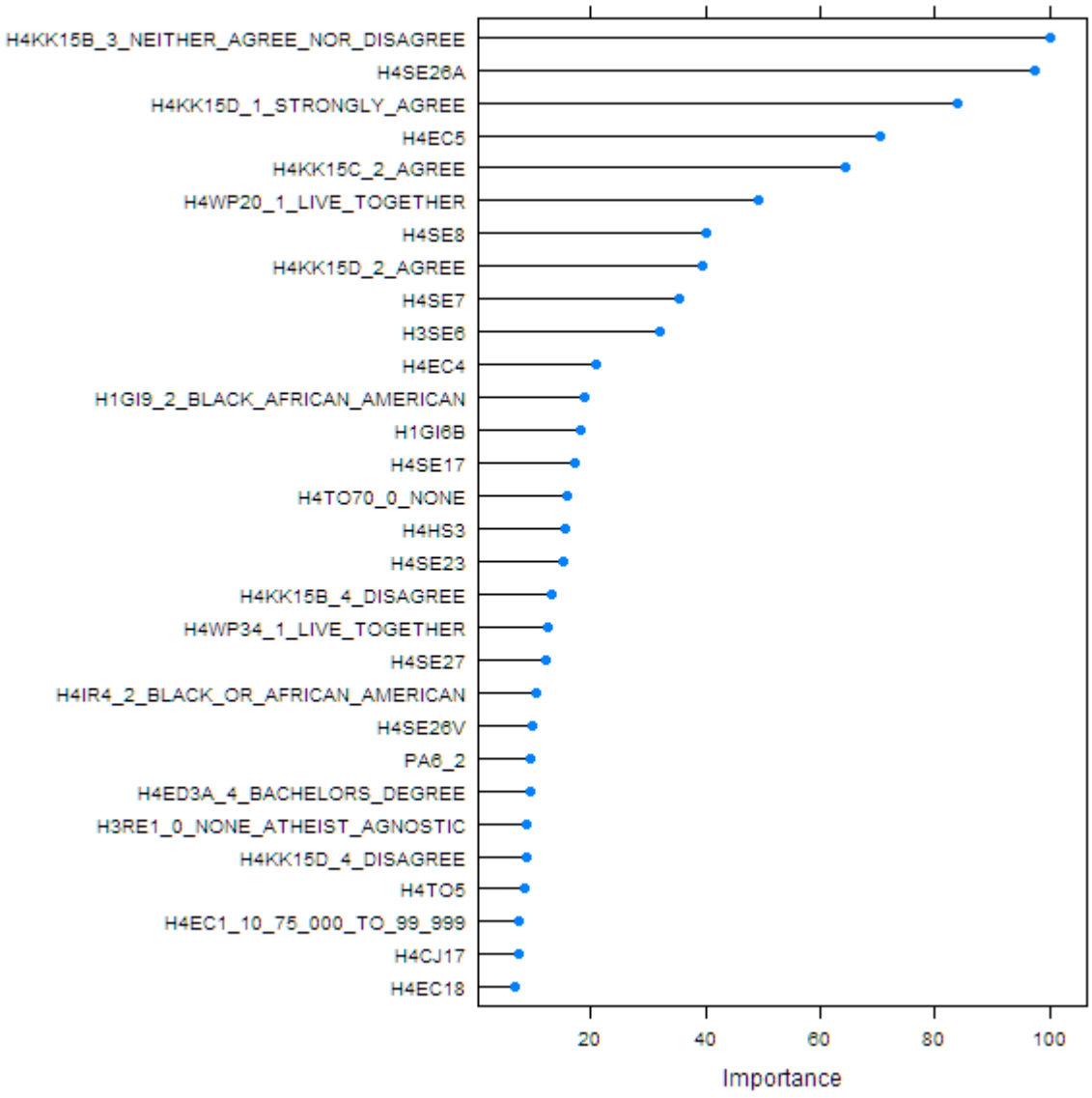


Table 4

Summary details of the top 30 variables of importance for the best model in Analysis II

Item	Item Text	Response	Wave	Code Book	Topics	Key Pattern	Importance	p	Sig.
H4KK15B	How much do you agree or disagree with the following statement? I feel close to my child(ren).	3 - neither agree nor disagree	4	Children and Parenting	Childrearing	When selecting this response: more TrueSingleton and PureCohabitation and less MarriedAtLeastOnce	100.00	0.18	
H4SE26A	In the past 12 months, did you or your partner(s) use any of these methods for birth control or disease prevention (check all that apply): condoms (rubbers)		4	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Contraception; Medications; Sexual Behavior	With yes response: more TrueSingleton and PureCohabitation and less MarriedAtLeastOnce	97.26	<.001	***
H4KK15D	How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being a parent.	1 - strongly agree	4	Children and Parenting	Childrearing; Stress/Anxiety	When selecting this response: more PureCohabitation and less MarriedAtLeastOnce	83.90	<.001	***
H4EC5	About how much do {YOU AND/OR YOUR SPOUSE/PARTNER} owe on the mortgage for your house, apartment, or residence?		4	Economics	Finances/SES; Household Characteristics; Marriage	With greater values: more MarriedAtLeastOnce and less PureCohabitation	70.35	0.03	*
H4KK15C	How much do you agree or disagree with the following statement? The major source of stress in my life is my child(ren).	2 - agree	4	Children and Parenting	Childrearing; Stress/Anxiety	When selecting this response: slightly more PureCohabitation and slightly less TrueSingleton and MarriedAtLeastOnce	64.23	0.39	
H4WP20	How far do you and your [mother figure] live from one another?	1 - live together	4	Parental Support and Relationships	Parental Support & Relationship	When selecting this response: more TrueSingleton and less MarriedAtLeastOnce	49.13	<.001	***
H4SE8	With how many partners have you ever had vaginal intercourse, even if only once?		4	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Sexual Behavior	With greater values: more PureCohabitation and less TrueSingleton and MarriedAtLeastOnce	39.93	<.001	***
H4KK15D	How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being a parent.	2 - agree	4	Children and Parenting	Childrearing; Stress/Anxiety	When selecting this response: more TrueSingleton and PureCohabitation and less MarriedAtLeastOnce	39.49	0.01	**
H4SE7	How old were you the first time you ever had vaginal intercourse?		4	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Sexual Behavior	With greater values: more TrueSingleton and less PureCohabitation	35.20	<.001	***
H3SE6	How many times have you had vaginal intercourse in the past 12 months?		3	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Sexual Behavior	With greater values: more PureCohabitation and MarriedAtLeastOnce and less TrueSingleton	31.99	<.001	***
H4EC4	Is your house, apartment, or residence owned or being bought by {YOU AND/OR YOUR SPOUSE/PARTNER}?		4	Economics	Finances/SES; Household Characteristics	With yes response: more MarriedAtLeastOnce and less PureCohabitation and TrueSingleton	21.04	<.001	***
H1GI9	Interviewer: Please code the race of the respondent from your observation alone.	2 - Black or African American	1	General Introductory	Race/Ethnicity	When selecting this response: more PureCohabitation and TrueSingleton and less MarriedAtLeastOnce	18.82	<.001	***

| Item | Item Text | Response | Wave | Code Book | Topics | Key Pattern | Importance | p | Sig. |

H1GI6B	What is your race (check all that apply): black or African American		1	General Introductory	Race/Ethnicity	With yes response: more PureCohabitation and TrueSingleton and less MarriedAtLeastOnce	18.05	<.001	***
H4SE17	Considering all types of sexual activity, with how many male partners have you had sex in the past 12 months, even if only one time?		4	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Sexual Behavior	With greater values: more TrueSingleton and PureCohabitation and less MarriedAtLeastOnce	17.01	0.15	
H4TO70	During the past 12 months, on how many days did you use marijuana?	0 - none	4	Tobacco, Alcohol, and Drugs	Illicit Drug Use; Marijuana; Substance Use/Abuse	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation	15.91	<.001	***
H4HS3	Over the past 12 months, how many months did you have health insurance?		4	Access to Health Services, Health Insurance	Health Insurance	With greater values: more MarriedAtLeastOnce and less PureCohabitation	15.37	<.001	***
H4SE23	Considering all types of sexual activity, with how many female partners have you had sex in the past 12 months?		4	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Sexual Behavior	With greater values: more PureCohabitation and less MarriedAtLeastOnce	15.01	<.001	***
H4KK15B	How much do you agree or disagree with the following statement? I feel close to my child(ren).	4 - disagree	4	Children and Parenting	Childrearing	When selecting this response: more PureCohabitation and less MarriedAtLeastOnce and TrueSingleton	13.16	0.35	
H4WP34	How far do you and your [father figure] live from one another?	1 - live together	4	Parental Support and Relationships	Parental Support & Relationship	When selecting this response: more TrueSingleton and less MarriedAtLeastOnce	12.35	<.001	***
H4SE27	In the past 12 months, did you have sex with more than one partner at around the same time?		4	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Risky sexual behavior; Sexual Behavior	With yes response: more PureCohabitation and TrueSingleton and less MarriedAtLeastOnce	12.11	<.001	***
H4IR4	Indicate the race of the sample member/respondent from your own observation (not from what the respondent said).	2 - Black or African American	4	Field Interviewer's Report	Race/Ethnicity	When selecting this response: more PureCohabitation and TrueSingleton and less MarriedAtLeastOnce	10.35	<.001	***
H4SE26V	In the past 12 months, did you or your partner(s) use any of these methods for birth control or disease prevention (check all that apply): no method used		4	Suicide, Sexual Experiences, and Sexually Transmitted Diseases	Contraception; Medications; Sexual Behavior	With yes response: more MarriedAtLeastOnce and less PureCohabitation and TrueSingleton	9.63	<.001	***
PA6_2	What is your race (check all that apply): Black/African American		1	Parent In-Home Index: Core	Parental Background Information; Race/Ethnicity	With yes response: more PureCohabitation and TrueSingleton and less MarriedAtLeastOnce	9.56	<.001	***
H4ED3A	Please list all degrees or certificates you have received from a college, university, or vocational/technical school. Do not include certificates you received from programs that lasted less than one year. What is the most recent degree you have received?	4 - bachelor's degree	4	Education	Attainment/Degrees/ Certificates	When selecting this response: more TrueSingleton and less PureCohabitation and MarriedAtLeastOnce	9.42	<.001	***
H3RE1	What is your present religion?	0 - none, atheist, agnostic	3	Religion and Spirituality	Religion & Spirituality	When selecting this response: more PureCohabitation and less MarriedAtLeastOnce	8.69	<.001	***

Item	Item Text	Response	Wave	Code Book	Topics	Key Pattern	Importance	p	Sig.

H4KK15D	How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being parent.	4 - disagree	4	Children and Parenting	Childrearing; Stress/Anxiety	When selecting this response: slightly more MarriedAtLeastOnce and slightly less PureCohabitation	8.59	0.24	
H4TO5	During the past 30 days, on how many days did you smoke cigarettes?		4	Tobacco, Alcohol, and Drugs	Substance Use/Abuse; Tobacco	With greater values: more PureCohabitation and less TrueSingleton and MarriedAtLeastOnce	8.42	<.001	***
H4EC1	Thinking about your income and the income of everyone who lives in your household and contributes to the household budget, what was the total household income before taxes and deductions in {2006/2007/2008}? Include all sources of income, including non-legal sources.	10 - $75,000 to $99,999	4	Economics	Finances/SES; Household Characteristics	When selecting this response: more MarriedAtLeastOnce and less PureCohabitation and TrueSingleton	7.47	<.001	***
H4CJ17	Have you ever spent time in a jail, prison, juvenile detention center or other correctional facility?		4	Involvement with Criminal Justice System	Incarceration	With yes response: slightly more PureCohabitation and less MarriedAtLeastOnce	7.26	0.21	
H4EC18	Between {1995/2002} and {2006/2007/2008}, did you or others in your household receive any public assistance, welfare payments, or food stamps?		4	Economics	Finances/SES; Household Characteristics	With yes response: more PureCohabitation and less TrueSingleton and MarriedAtLeastOnce	6.80	<.001	***

Results Discussion for Analysis II

In Analysis II, lifestyle choice at Wave IV was predicted using variables from Waves I through IV. Overall, a gradient boosted model was the most accurate, correctly predicting test cases nearly 75% of the time. As shown in the variables of importance in Table 4, this model incorporated a variety of interesting independent variables that warrant further discussion. Because many of the variables were similar to one another, it is possible to categorize them based on their similarity. To achieve this, the code book from which each item was drawn along with the topic level groupings provided in the Add Health Codebook Explorer (2017) were taken into account. The resulting items category sets are further discussed below and tied into existing literature on marriage, cohabitation, and singlehood. More specifically, the categories are elaborated upon sequentially based on the overall importance of the top item in a given set relative to the variables of importance shown in Table 4 above. Table 6 below provides a summary of the item sets as they will be presented. Given the commonalities between Analysis I

and Analysis II, the following section will provide an overview of the categories already discussed from Analysis I above and discuss the unique categories in detail.

Table 6
Analysis II: Categorization of the top variables of importance from Table 4

Category	Unique Items	Items
Children and Parenting	3	H4KK15B, **H4KK15D**, H4KK15C
Sexual Experiences, STDs, and Health	8	**H4SE26A**, H4SE8, H4SE7, H3SE6, H4SE17, H4SE23, H4SE27, H4SE26V
Economics and Employment	4	H4EC5, H4EC4, H4EC1, **H4EC18**
Parents	3	**H4WP20**, H4WP34, PA6_2
Race/Ethnicity	3	H1GI9, **H1GI6B**, H4IR4
Tobacco and Marijuana	2	**H4TO70**, H4TO5
Access to Health Services/Insurance	1	**H4HS3**
Education	1	**H4ED3A**
Religion and Spirituality	1	**H3RE1**
Involvement with Criminal Justice System	1	**H4CJ17**

Note: Children and Parenting contains two recurring items, and therefore the total number of unique items is 28 rather than 30. The associated plots for the items in **bold** are included in the sections below. All other plots may be found in Appendix D. Categories in *italics* are also included in Analysis I above.

Overview of Repeated Categories

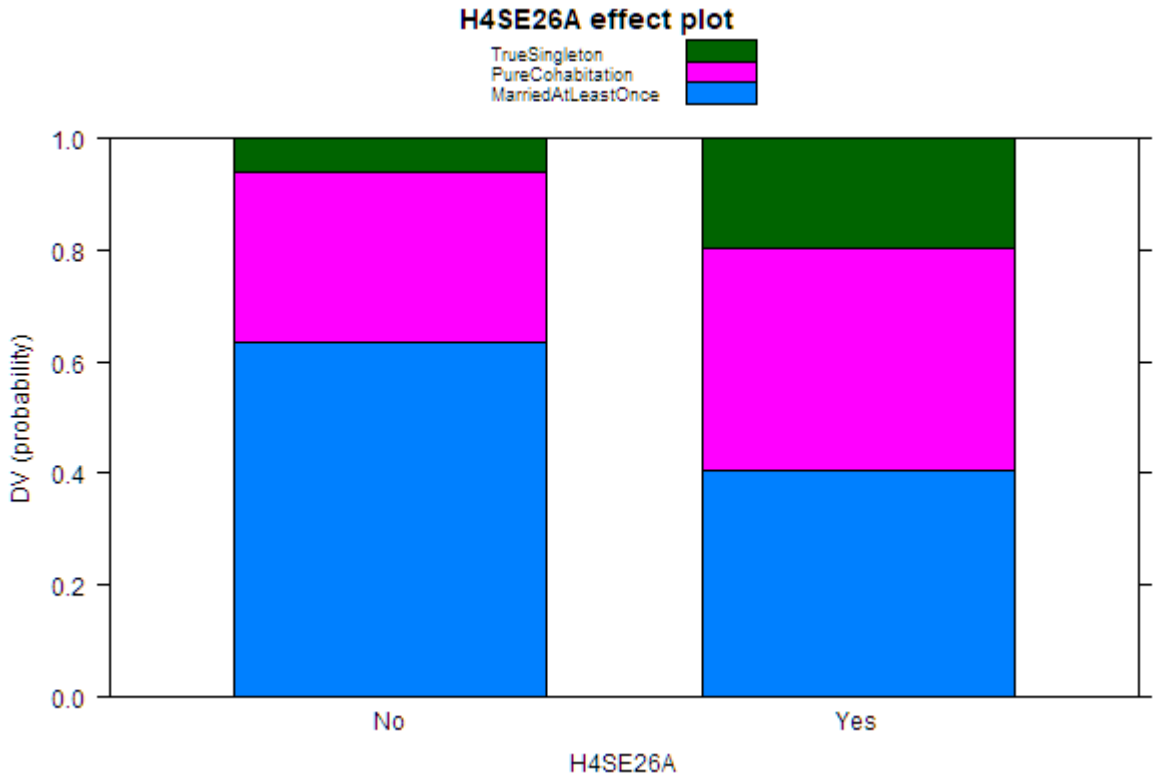
Analysis II found many similar groupings as Analysis I: Sexual Experiences, STDs, and Health; Economics and Employment; Parents; Race/Ethnicity; Tobacco and Marijuana; Education; and Religion and Spirituality. The overarching subject and content within each category remain aligned with Analysis I; however, due to the inclusion of Wave IV data in Analysis II, the individual items that were found to be of importance vary slightly.

Sexual Experiences, STDs, and Health

This topic grouping - Sexual Behavior, including risky sexual behavior, STDs, and

contraception use - contains eight distinct items. They are listed here in descending magnitude: H4SE26A (see Figure 22 below), H4SE8, H4SE7, H3SE6, H4SE17, H4SE23, H4SE27, and H4SE26V. The actual text of the items and their effects are summarized in Table 4 above.

Figure 22
H4SE26A

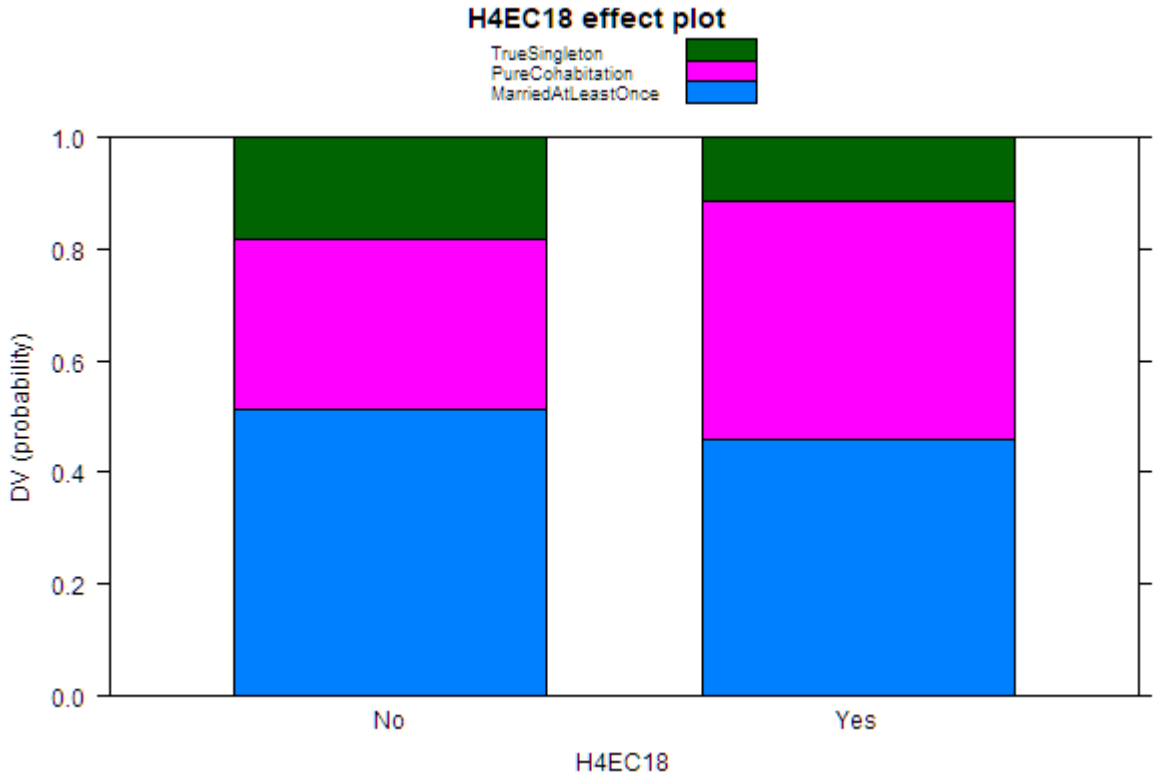


Note. For this item, the Yes selection indicates that they did not use any method of contraception or disease prevention based on the following item: "In the past 12 months, did you or your partner(s) use any of these methods for birth control or disease prevention (check all that apply): no method used." In other words, individuals who indicated they did not use birth control or disease prevention were more likely to be singletons or cohabitors and less likely to be married.

Economics and Employment

This category contains four items: H4EC5, H4EC4, H4EC1, and H3EC18 (see Figure 23 below). These items collectively tap into homeownership and accumulated wealth, and the general pattern is that both greater wealth and homeownership are associated with individuals being more likely to be married.

Figure 23
H3EC18

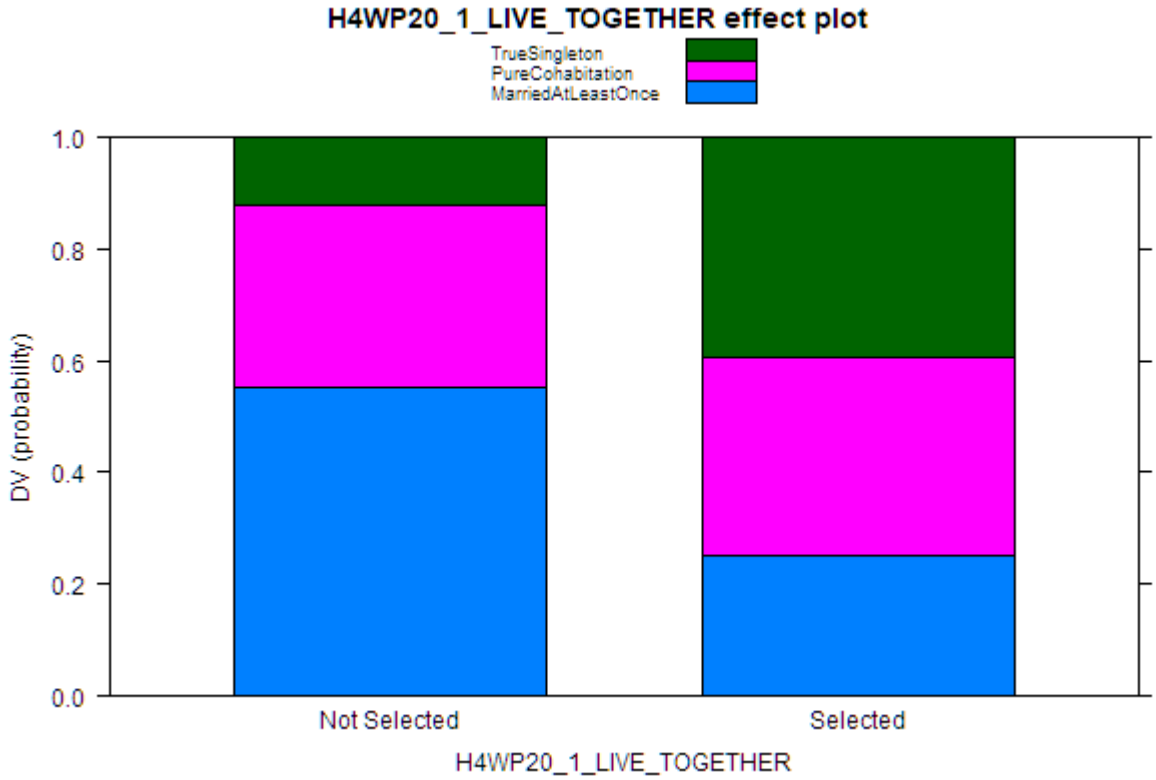


Note. This item asked: "Between {1995/2002} and {2006/2007/2008}, did you or others in your household receive any public assistance, welfare payments, or food stamps?" Selecting Yes indicates that the participant received some type of public assistant between the given years. An individual who did not receive any public assistance (i.e., selecting No), was more likely to be married.

Parents

The next item set - Parents - contains three items: H4WP20 (see Figure 24 below), H4WP34, and PA6_2. More broadly, all of these items relate to the degree of involvement the adult child has with their mother or father, and the former two are highly significant predictors. In general, people still living with their parents are not at the lifespan developmental point at which most people seek a partnered relationship either in marriage or cohabitation.

Figure 24
H4WP20



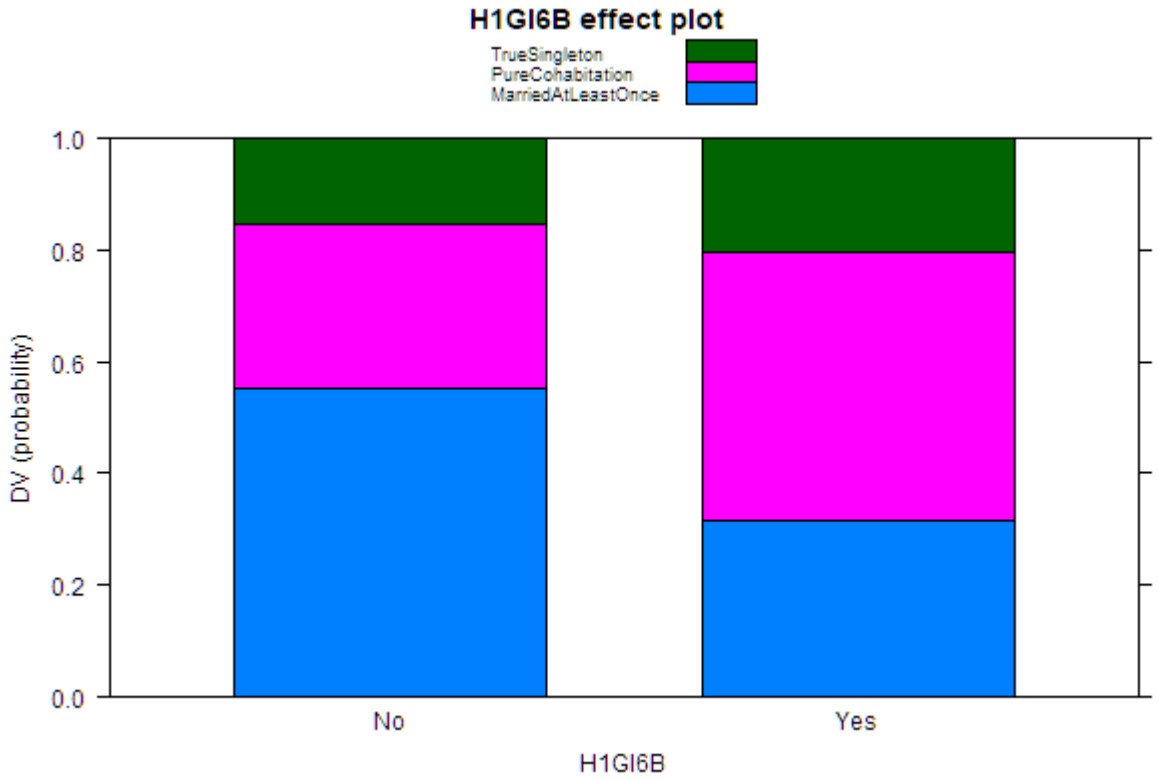
Note. "How far do you and your [mother figure] live from one another?" For this item, Selected indicates the participant chose the response Live Together. Living with one's mother figure showed a high probability of singlehood, while not selecting Live Together was highly predictive of marriage.

Race/Ethnicity

This topic grouping contains one item related to the respondents self-identification as Black or African American, H1GI6B (see Figure 25 below), and two items related to race/ethnicity as perceived by the interviewer: H4IR4 and H1GI9, which ask whether an individual is African American or White, respectively. These three items show that White individuals and non-Black or non-African American individuals are more likely to be married, whereas non-White individuals are more likely to be cohabitors or singletons, which is particularly the case for African-Americans (i.e., based on H1GI6B).

Figure 25

H1GI6B

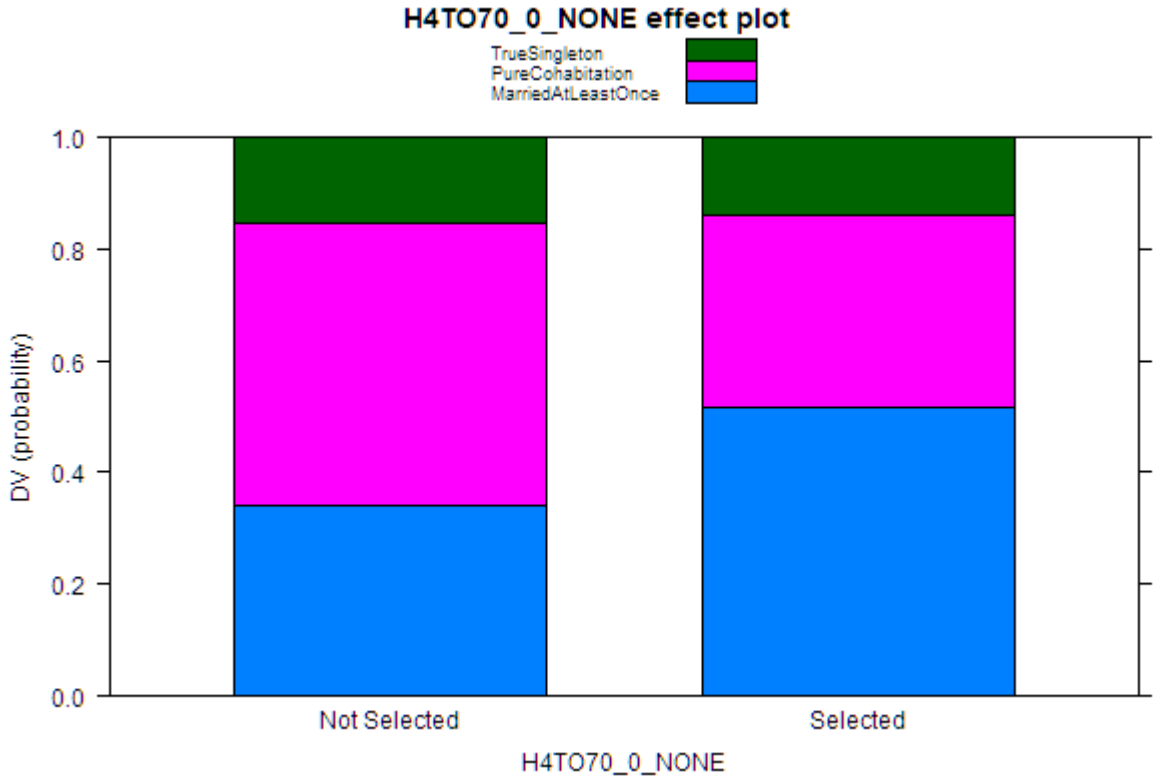


Note. H1GI9: For the item, "Interviewer: Please code the race of the respondent from your observation alone," Yes indicates the respondent selected Black or African American as their race. No had a higher probability of marriage, whereas Yes had a higher probability of cohabitation.

Tobacco and Marijuana

This topic set contains two items, H4TO70 (see Figure 26 below) and H4TO5. H4TO70 showed that individuals who indicated they did not use marijuana in the past 12 months were more likely to be married than to cohabitate, with little to no change for singletons.

Figure 26
H4TO70



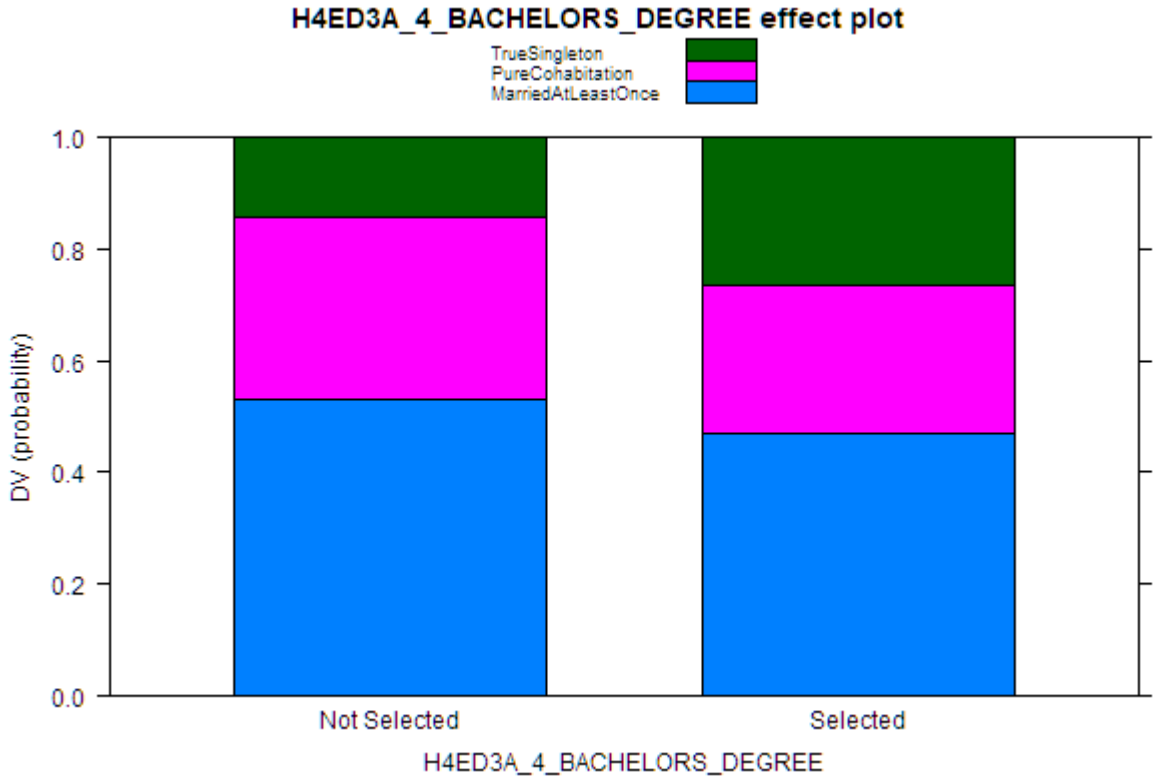
Note. For the item, "During the past 12 months, on how many days did you use marijuana?," Selected indicates the individual selected they used marijuana zero days in the past 12 months. Individuals who used marijuana zero days in the past 12 months were more likely to be married than the Not Selected respondents.

Education

This topic contains a single education-related item, H4ED3A (see Figure 27 below), which determines if a participant's most recent degree was a bachelor's. Individuals who attained this degree were more likely to be married and less likely to be cohabitors on average. This finding is consistent with research that suggests that educational attainment plays a significant role in perceived attractiveness of marital partners and subsequent marriage transition (Stevens et al., 1990). More recent research suggests that education is a driver in the marriage divide whereby a smaller number of highly educated individuals intermarry at greater rates whereas those with lower levels of education are more likely to remain cohabitors or stay single (Scwartz

& Mare, 2005).

Figure 27
H4ED3A

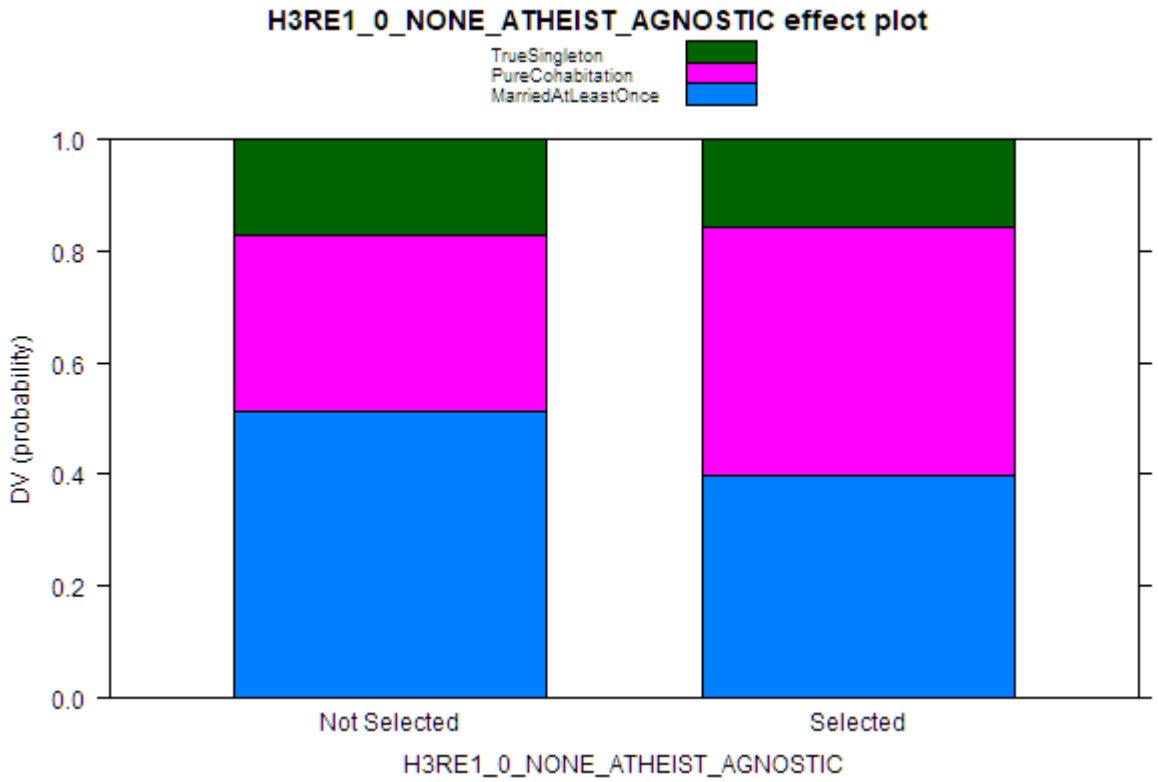


Note. This item states, "Please list all degrees or certificates you have received from a college, university, or vocational/technical school. Do not include certificates you received from programs that lasted less than one year. What is the most recent degree you have received?" Selected indicated the respondent chose Bachelor's Degree as the most recent degree received. These individuals were more likely to be single, whereas individuals in the Not Selected group were more likely to be married.

Religion and Spirituality

This topic contains a single item on religious affiliation from Wave III, H3RE1 (see Figure 28 below): "What is your present religion?" Individuals who indicated they had no religious affiliation or were atheist or agnostic were less likely to be married and more likely to cohabitate. Those who endorsed any religion were more likely to be married.

Figure 28
H3RE1

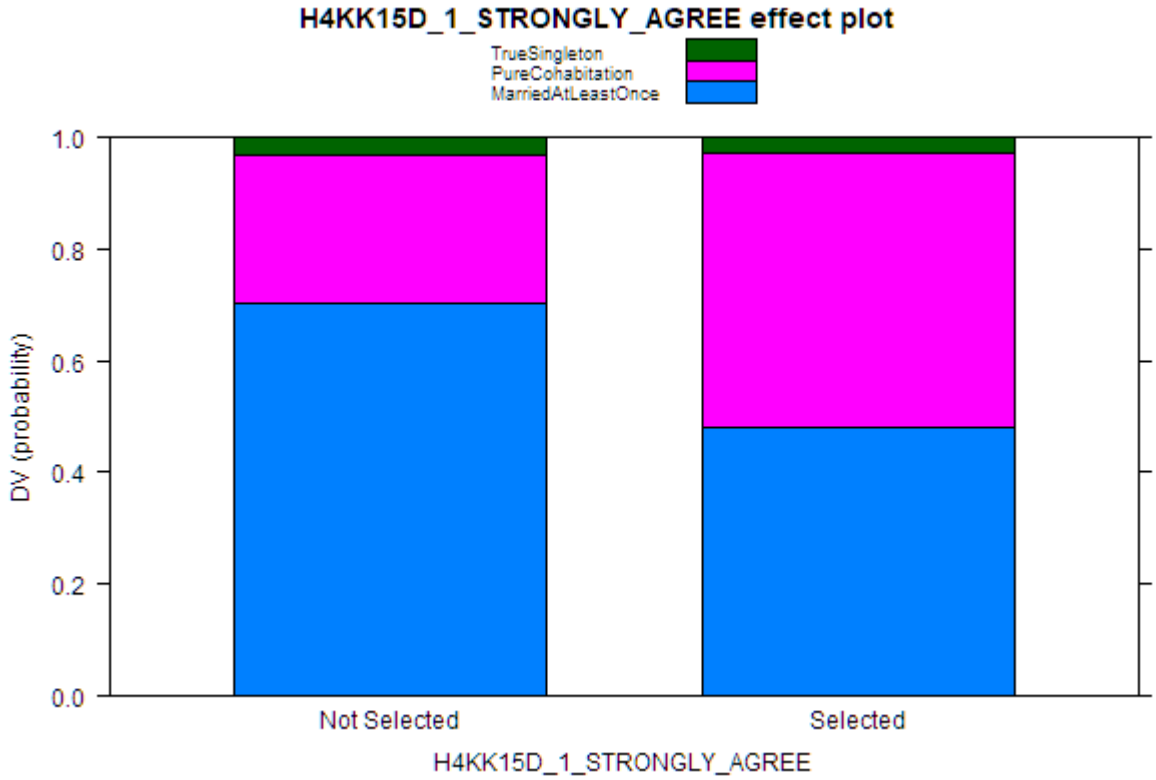


Note. This item states, "What is your present religion?" Selected indicates the respondent chose None, Atheist, Agnostic. Not Selected shows a higher probability of marriage, whereas Selected shows a fairly equal likelihood of marriage and cohabitation, with a slightly higher probability of cohabitation over singlehood.

Children and Parenting

The first unique topic grouping (not included in Analysis I) - Children and Parenting - contains three unique items, with three that recurred in the variables of importance (i.e., where the model used multiple levels of a single originating factor variable). The three unique items in this set relative to their descending order of importance are H4KK15B: "I feel close to my child(ren);" H4KK15D: "I feel overwhelmed by the responsibility of being a parent" (see Figure 29 below); and H4KK15C: "The major source of stress in my life is my child(ren)" (Add Health Codebook Explorer, 2017). All three items were binary and related to specific responses on the original scale of agreement (i.e., "How much do you agree or disagree with the following statement?" ranging from strongly disagree to strongly agree on a five-point Likert scale).

Figure 29
H4KK15D



Note. For this item, the Not Selected response indicates the participant did not select "strongly agree" for the item, "How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being a parent?" Selected indicates that the participant did select that they strongly agree with the item. The plot demonstrates that individuals who strongly agree with feeling overwhelmed by the responsibility of being a parent are less likely to be married and more likely to cohabit than those who did not select "strongly agree."

As shown in the effect summaries presented in Table 6, people who experience greater stress around parenting are more likely to be cohabitors, whereas those who say the opposite is more likely to be married. With regard to marriage, there are connections between engagement in childrearing and perceived marital stability (Kalmijn, 1999); parental investment, marriage promoting behaviors, and benefits experienced by children (Schultz, 1974); and developmental perspectives on the interaction between childrearing attitudes their effects on marital quality and subsequent child development (Gable et al., 1992). Each of these studies showed a connection between attitudes about parenting and the decision to marry. Regarding cohabitation, research suggests that attitudes regarding parenting and cohabitation have been changing, with

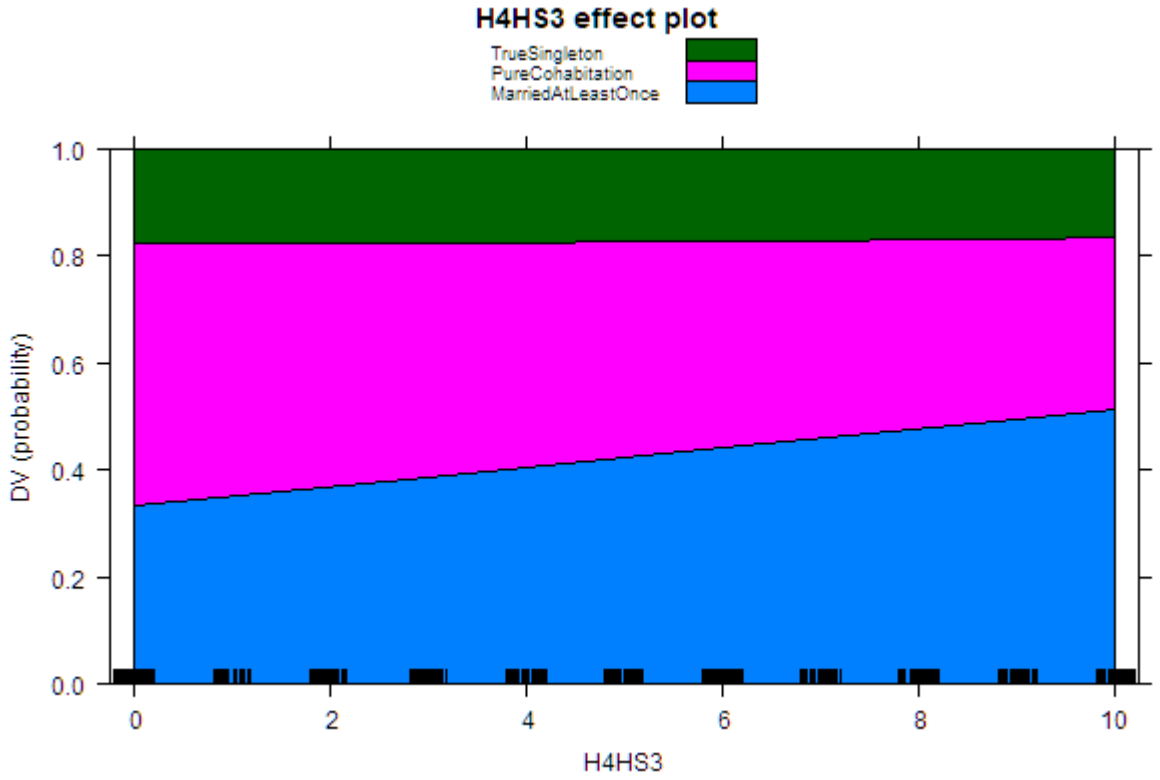
cohabitation becoming more commonplace (Timberlake & Heuveline, 2005). The stress of childrearing may be felt more intensely by cohabitors because they cannot rely on the stability provided by a lawful marriage. Beyond this, unfortunately, research on cohabitation and childrearing is relatively sparse.

Access to Health Services/Insurance

The second unique topic has only one item which asks about access to health insurance over the past 12 months (H4HS3; see Figure 30 below). On the one hand, this variable may appear simply as a proxy for economic stability. However, as Waldron et al. (1996) point out, it may be the case that healthy people are more likely to get married, and disentangling the health benefits of marriage from the process of healthier individuals choosing to marry at greater rates is a complex issue. This is especially true in the US, where access to health care is economically

stratified, making it particularly challenging to draw definitive conclusions (Jemal et al., 2008).

Figure 30
H4HS3

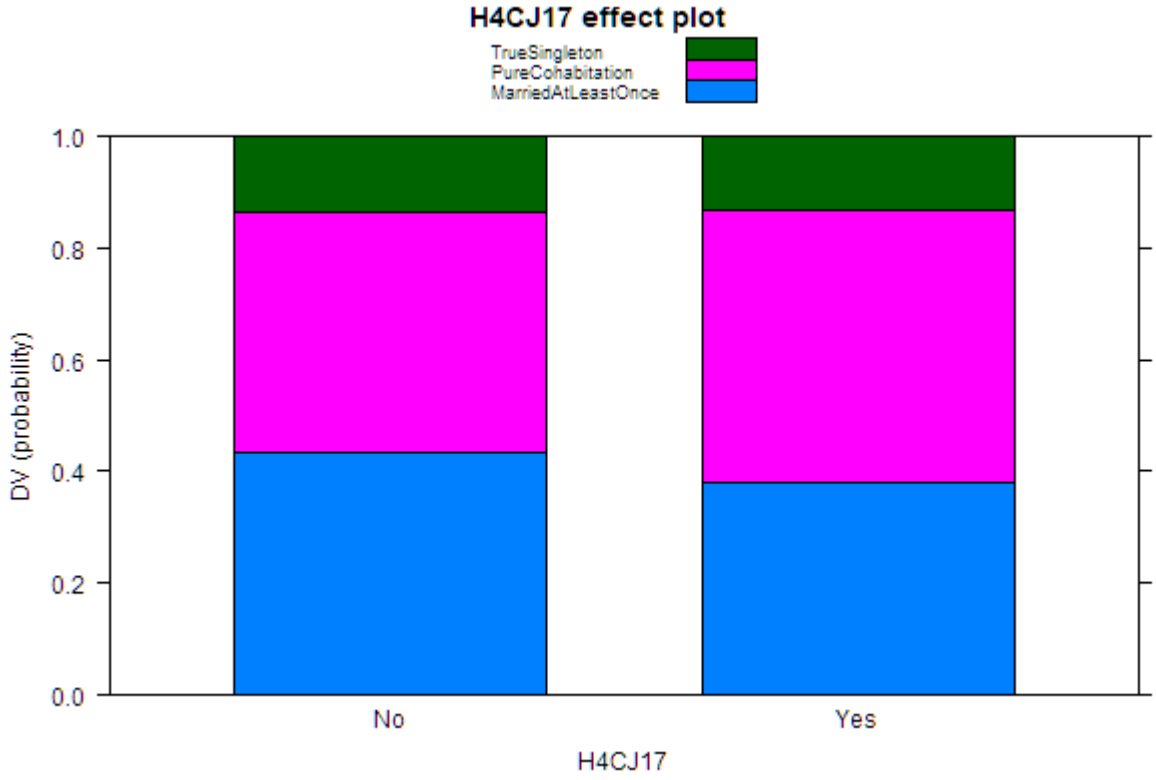


Note. This is a continuous item: "Over the past 12 months, how many months did you have health insurance?" The probability of singlehood remains mostly consistent across selections (0 months to 10 months), whereas the likelihood of marriage increases as the total months of coverage increases (e.g., an individual who had at least 10 months of coverage is more likely to be married than an individual who had 0 months of coverage).

Involvement with Criminal Justice System

The last unique category related to involvement in the criminal justice system and contained one item: H4CJ17 (see Figure 31 below). This item has to do with whether or not an individual has been incarcerated. Individuals with greater involvement with the criminal justice system were more likely to be cohabitors. In terms of this finding, research has shown that men who experience incarceration are subject to a number of challenges through employment difficulty, discrimination and reduced social capital that leads to decreased marriage prospects compared to non-incarcerated individuals, contributing to the pattern of lifestyle choice demographics in America (Western & Lopoo, 2004). Similarly, another recent study indicated that incarceration had a direct negative effect with entry to marriage that did not similarly exist for entry into cohabitation (Apel, 2016).

Figure 31
H4CJ17



Note. For this item, "Have you ever spent time in a jail, prison, juvenile detention center or other correctional

facility?," Yes indicates the individual has spent time in a jail, prison, juvenile detention center, or other correctional facility. Individuals who selected Yes were more likely to cohabit, whereas there was essentially no change between No and Yes for singlehood.

General Discussion

Existing research on the predictive factors associated with an individual's lifestyle choice (i.e., marriage, cohabitation, and singlehood) has led to a wide variety of interesting findings. Unfortunately, given the multitude of variables that can influence an individual's lifestyle choice, these findings have often been inconsistent or even contradictory. To resolve this problem, it is clear that an approach that is able to incorporate vast amounts of data is the optimal pathway for analyses that encompass the full breadth of individual differences. To that end, the machine learning models constructed in this study were trained on as large a set of variables as possible in a longitudinal dataset to uncover the variables most likely to predict an individual's lifestyle choice.

As established in the Literature Review, research has shown that a wide range of benefits accompanies marriage and cohabitation, including increased life satisfaction, improved overall mental, behavioral, and medical health, elevated levels of education and socioeconomic status, and more. In fact, one of the key important predictors in this study showed that individuals who are very satisfied with their lives are more likely to be married (i.e., H3SP3: "How satisfied are you with your life as a whole?"). As such, marriage and cohabitation are two lifestyle choices that are, in general, beneficial to individual well-being as well as to societal well-being. Specifically, there are some clear benefits (e.g., behavioral, health, and medical) of having a supportive partner, be it from marriage or cohabitation. This has been demonstrated through various studies (e.g., Hayes et al., 2016), as well as through this study's analyses. There is a greater likelihood that married individuals will experience positive economic and household benefits due to the legal status associated with marriage. However, cohabitors will also

experience many of the same positive benefits due to the protective effects of having a supportive partner, including improved mental health.

Consistent with prior work, the analyses presented in this study showed that these benefits might be dampened for certain demographics, including race, socioeconomic status (SES), and education. Moreover, singletons are less likely to experience these benefits: rather, they are more likely to suffer from single strain and have a higher likelihood of health complications later in life.

These analyses identified several interesting variables that may point to important behavioral health issues within America. As mentioned in Table 5 and Table 6, these variables were grouped into categories based on similarities. When reviewing these categories and their associated effect plots, certain variables and categories stand out. Specifically, Sexual Experiences are highly predictive of singlehood; Religion and Spirituality are highly predictive of singlehood or marriage (very religious individuals are far less likely to cohabit); and Race/Ethnicity, Economics and Employment, and Education are all indicative of cohabitation and marriage. Poorly educated Black or African American individuals are more likely to cohabit than their non-Black or non-African American counterparts.

Based on the existing literature, it is evident that things like education, money, and race are highly stratified in America. Such stratification greatly impacts the overall mental and physical well-being of individuals. The relationships identified within these data suggest that individuals who are poorly educated, low SES, and/or come from disruptive family systems are more likely to be less satisfied with their life as a whole, experience greater levels of depression, and engage in detrimental health behaviors. It is important to note that these correlations result from societal stratification, systematic biases, and racial oppression. It is not clear if these

variables can be disentangled from this analysis; however, the findings from this study provide an overlay for further, more traditional research.

Limitations and Future Research Directions

The machine learning approach used in this study yielded an informative ranking of key variables that predict marriage, cohabitation, and singlehood in a unified model of Waves I through IV in the Add Health dataset. While a variety of studies have been conducted using the Add Health dataset previously, this study incorporated a significantly larger set of variables to be as comprehensive as possible. Based on the results presented and the variables of importance uncovered, it should now be possible for experts in their respective fields to pursue specific results of interest in future follow-up studies using more traditional statistical methods. These follow-up analyses are particularly important since the results of these analyses are based on observational data and should not be interpreted causally.

More broadly, the general approach outlined in this dissertation could be adapted to explore other dependent variables in Add Health (e.g., medical or mental health outcomes). Alternatively, this approach could also be used to analyze other large, longitudinal datasets to produce similarly useful, structured results to guide future research across a variety of domains.

To be clear, this researcher does not view machine learning as a replacement for standardized methods but as a tool for efficiently prioritizing variables for future studies in situations where items and their associated observations are exceptionally large and complex (e.g., large sets of items with complex interrelationships and or longitudinal/data with complex survey structures across multiple waves). It is strongly suggested that follow-up studies be conducted to investigate key patterns found in this research in a more nuanced fashion. More broadly, an exploratory finding of this analysis revealed some complicated relationships among

these predictive factors. Future research should analyze these data in order to establish interventions aimed at encouraging marriage and/or cohabitation and educating individuals.

Finally, while it was unavailable while conducting this dissertation research, the new Add Health Wave V dataset has recently been made available to researchers upon request, and it would likely be fruitful to include it in a future machine learning analysis that builds upon the work presented here. Doing so would help to provide a more comprehensive picture of marriage, cohabitation, and singlehood status into later life stages since the current study is limited to ages 24 to 32 (i.e., the age range captured by Wave IV).

Conclusion

Throughout this dissertation, the complexity of lifestyle choice in a large and diverse sample (i.e., the Add Health dataset) was examined to predict three distinct outcomes: marriage, cohabitation, and singlehood. Examining these three relationship outcomes collectively was necessary due to the greater diversity of relationship choices that have materialized with the US' changing demographic patterns in recent decades. In the past, traditional research has focused on studying lifestyle choice by using hypothesis-tests and selecting a handful of variables. While these studies have undoubtedly been fruitful, exploratory machine learning (ML) approaches offer a novel way to uncover robust and highly explanatory predictors in models that encompass the full complexity of large, complex, longitudinal datasets such as Add Health. It should be expected that specific groups will have unique pathways to a particular lifestyle choice and that none of these groups should go bereft of empirical scrutiny. Machine learning models are uniquely capable of uncovering such findings.

Critics of ML approaches may argue that the examination of rich, longitudinal data will only reveal the most broadly applicable and blunt findings. Instead, the variables of importance

generated by these analyses highlighted several unique factors relevant to diverse groups (e.g., marijuana) in addition to seemingly canonical predictors such as income and education - both types of variables ranked highly in my best model, which was over 75% accurate overall in predicting lifestyle choice. Ultimately, this researcher believes these analyses - and the general framework presented - will satisfy the desires of empiricists who hope to find differences that matter (i.e., those with substantial effect sizes) as well as researchers in search of more nuanced and novel patterns particular to specific subsets of the population.

References

Add Health Codebook Explorer (ACE). (2017, June 20). Retrieved from https://www.cpc.unc.edu/projects/addhealth/documentation/ace

Allen, K. R. (1994). Feminist reflections on lifelong single women. *Gender, Families and Close Relationships: Feminist Research Journeys*, *2*, 97-105. https://books.google.com/books?hl=en&lr=&id=L7N1AwAAQBAJ&oi=fnd&pg=PA97&dq=Feminist+reflections+on+lifelong+single+women&ots=-keqNCbrkK&sig=N8CiNpi3v5mgIMtuWddGd-Khho8#v=onepage&q=Feminist%20reflections%20on%20lifelong%20single%20women&f=false

Allen, E. S., Rhoades, G. K., Stanley, S. M., Loew, B., & Markman, H. J. (2012). The effects of marriage education for army couples with a history of infidelity. *Journal of Family Psychology*, *26*(1), 26.

Amato, P. R. (2015). Marriage, cohabitation and mental health. *Family Matters*, (96), 5.

Apel, R. (2016). The effects of jail and prison confinement on cohabitation and marriage. *The ANNALS of the American Academy of Political and Social Science*, *665*(1), 103-126.

Axinn, W. G. & Thornton, A. (1993). Mothers, children, and cohabitation: The intergenerational effects of attitudes and behavior. *American Sociological Review*, 233-246.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research, 20*(1), 40-49.

Band-Winterstein, T. & Manchik-Rimon, C. (2014). The experience of being an old never married single: A life course perspective. *The International Journal of Aging and Human Development, 78*(4), 379-401.

Billari, F. C., Fürnkranz, J., & Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population/Revue Européenne de Démographie, 22*(1), 37-65.

Bellani, D., Esping-Anderson, G. &, Lesia, N. (2017). Never partnered: A multilevel analysis of lifelong singlehood. *Demographic Research, 37*(4), 53-100. 10.4054/DemRes.2017.37.4

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making, 16*(3), 74.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13,* 281-305.

Black, D. A., Heyman, R. E., & Slep, A. M. S. (2001). Risk factors for child physical abuse. *Aggression and Violent Behavior, 6(*2-3), 121-188.

Booth, A. & Johnson, D. (1988). Premarital cohabitation and marital success. *Journal of Family Issues, 9*(2), 255-272.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Brein, M. J., Lillard, L. A., & Stern, S. (2006). Cohabitation, marriage, and divorce in a model of match quality. *International Economic Review, 47*(2), 451-494.

Brown, S. L., Lee, G. R., & Bulanda, J. R. (2006). Cohabitation among older adults: A national portrait. *The Journals of Gerontology, 61*B(2), S71-S79.

Bryner, J. (2011). *Close friends less common today, study finds.* LiveScience. https://www.livescience.com/16879-close-friends-decrease-today.html

Bumpass, L. L. & Sweet, J. A. (1989). National estimates of cohabitation. *Demography, 26*(4), 615-625.

Bumpass, L. L., Sweet, J. A., & Cherlin, A. (1991). The role of cohabitation in declining rates of marriage. *Journal of Marriage and the Family,* 913-927.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121-167.

Busby, D. M., Carroll, J. S., & Willoughby, B. J. (2010). Compatibility or restraint? The effects of sexual timing on marriage relationships. *Journal of Family Psychology*, *24*(6), 766.

Caucutt, E. M., Guner, N., & Rauh, C. (2018). Is marriage for white people? Incarceration, unemployment, and the racial marriage divide.

CDC. (2017). *National marriage and divorce rate trends*. [Dataset]. CDC/NCHS National Vital Statistics System. https://www.cdc.gov/nchs/data/dvs/national-marriage-divorce-rates-00-18.pdf

CDC. (2020). Marriage and Divorce. https://www.cdc.gov/nchs/fastats/marriage-divorce.htm

Cherlin, A. J., Ribar, D. C., Yasutake, S. (2016). Nonmarital first births, marriage, and income inequality. *American Sociological Review, 81*(4), 749-770.

Chiappori, P. A., Oreffice, S., & Quintana-Domeque, C. (2010). Matching with a handicap: The case of smoking in the marriage market.

Chin, B., Murphy, M. L. M., Janicki-Deverts D, & Cohen, S. (2019). Marital status as a predictor of diurnal salivary cortisol levels and slopes in a community sample of healthy adults. *Psychoneuroendocrinology*, 78. 68–75.

Cohan, C. L. & Kleinbaum, S. (2002). Toward a greater understanding of the cohabitation effect: Premarital cohabitation and marital communication. *Journal of Marriage and Family*, *64*(1), 180-192.

Cohn, D'Vera. (2011). *Marriage rate declines and marriage age rises*. Pew Research Center. https://www.pewsocialtrends.org/2011/12/14/marriage-rate-declines-and-marriage-age-rises/

Cunningham, M. & Thornton, A. (2006). The influence of parents' marital quality on adult children's attitudes toward marriage and its alternatives: Main and moderating effects. *Demography*, *43*(4), 659–672.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-2792.

DeLap, H. (2000). Personal readiness for marriage in adult children of alcoholics and adult children of non-alcoholics. The Graduate College of Wisconsin-Stout.

DeMaris, A. & Rao, K. V. (1992). Premarital cohabitation and subsequent marital stability in the United States: A reassessment. *Journal of Marriage and the Family*, 178-190.

De Neve, J. E., Diener, E., Tay, L., & Xuereb, C. (2013). The objective benefits of subjective well-being. *World happiness report*.

DePaulo, B. M. & Morris, W. L. (2006). The unrecognized stereotyping and discrimination against singles. *Current Directions in Psychological Science, 15*(5), 251-254.

DePaulo, B. M., & Morris, W. L. (2005a). Should singles and the scholars who study them make their mark or stay in their place? *Psychological Inquiry, 16*, 142-149.

DePaulo, B. M., & Morris, W. L. (2005b). Singles in society and in science. *Psychological Inquiry*, *16*(2-3), 57-83.

DiClemente, R. J., Wingood, G. M., Crosby, R., Sionean, C., Cobb, B. K., Harrington, K., ... & Oh, M. K. (2001). Parental monitoring: Association with adolescents' risk behaviors. *Pediatrics*, *107*(6), 1363-1368.

Edin, K. & Reed, J. (2005). Why don't they just get married? Barriers to marriage among the disadvantaged. *The Future of Children, 15*(2), 117-137.

Eickmeyer, K. J. & Manning, W. D. (2018). Serial cohabitation in young adulthood: Baby boomers to millennials. *Journal of Marriage and Family, 80*(4), 826-840.

Eliason, S. R., Mortimer, J. T., & Vuolo, M. (2015). The transition to adulthood: Life course structures and subjective perceptions. *Social psychology quarterly*, *78*(3), 205-227.

Eryigit, S., Cadely, H. S., & Harrell-Levy, M. K. (2010). What adolescents bring to and learn from relationship education classes: Does social address matter? *Journal of Couple & Relationship Therapy, 9*, 95–112.

Esposito, M. H., Lee, H., Hicken, M. T., Porter, L. C., & Herting, J. R. (2017). The consequences of contact with the criminal justice system for health in the transition to adulthood. *Longitudinal and Life Course Studies, 8*(1), 57-74.

Fincham, F. D., & Beach, S. R. (2010). Marriage in the new millennium: A decade in review. J*ournal of Marriage and Family, 72(*3), 630-649.

Fokkema, T. & Liefbroer, A. C. (2008). Trends in living arrangements in Europe: Convergence or divergence? *Demographic Research, 19*(35), 1351-1418. 10.4054/DemRes.2008.19.36

Forste, R. & Tanfer, K. (1996). Sexual exclusivity among dating, cohabiting, and married women. *Journal of Marriage and the Family*, 33-47.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Japanese Society For Artificial Intelligence, 14*(771-780), 1612.

Fry, R. (2012). *No reversal in decline of marriage.* Pew Research Center. Washington, DC http://www. pewsocialtrends. org/2012/11/20/no-reversal-in-decline-ofmarriage

Gable, S., Belsky, J. & Crnic, K. (1992). Marriage, parenting, and child development: Progress and prospects. Journal of Family Psychology, 5(3-4), 276.

Garrison, M. (2007). The decline of formal marriage: Inevitable or reversible? *Family Law Quarterly, 41*(3), 491-520.

Goldin, C., & Katz, L. F. (2002). The power of the pill: Oral contraceptives and women's career and marriage decisions. *Journal of political Economy*, *110*(4), 730-770.

Grace, K., & Sweeney, S. (2014). Pathways to marriage and cohabitation in Central America. *Demographic Research, 30*, 187-226.

Gray, J. D. & Silver, R. C. (1990). Opposite sides of the same coin: Former spouses' divergent perspectives in coping with their divorce. *Journal of Personality and Social Psychology*, *59*(6), 1180.

Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers (2019). GBM: Generalized boosted regression models. R package version 2.1.5. Retrieved from https://CRAN.R-project.org/package=gbm

Greenwood, J., & Guner, N. (2008). Marriage and divorce since World War II: Analyzing the role of technological progress on the formation of households. *NBER Macroeconomics Annual, 23*(1), 231-276.

Gubernskaya, Z. (2010). Changing attitudes toward marriage and children in six countries. *Sociological Perspectives*, *53*(2), 179-200.

Gurney, J. G., Krull, K. R., Kadan-Lottick, N., Nicholson, H. S., Nathan, P. C., Zebrack, B., ... & Ness, K. K. (2009). Social outcomes in the childhood cancer survivor study cohort. *Journal of clinical oncology*, *27*(14), 2390.

Gurrentz, B. (2018). *Living with an unmarried partner now common for young adults.* US Census Bureau. https://www.census.gov/library/stories/2018/11/cohabitaiton-is-up-marriage-is-down-for-young-adults.html

Gurrentz, B. (2019). *Cohabiting partners older, more racially diverse, more educated, higher earners.* US Census Bureau. https://www.census.gov/library/stories/2019/09/unmarried-partners-more-diverse-than-20-years-ago.html

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Hacker, D. (2001). Single and married women in the law of Israel–A feminist perspective. *Feminist Legal Studies, 9*(1), 29-56.

Hall, S. S. (2006). Parental predictors of young adults' belief systems of marriage. *Current Research in Social Psychology*, *12*(2), 22-37.

Halpern, C. T., Waller, M. W., Spriggs, A., & Hallfors, D. D. (2006). Adolescent predictors of emerging adult sexual patterns. *Journal of Adolescent Health*, *39*(6), 926-e1.

Harakeh, Z., Scholte, R. H., Vermulst, A. A., de Vries, H., & Engels, R. C. (2004). Parental factors and adolescents' smoking behavior: An extension of the theory of planned behavior. *Preventive medicine*, *39*(5), 951-961.

Hawkins, A. J. (2013). The forever initiative: A feasible public policy agenda to help couples form and sustain healthy marriages and relationships. North Charleston, South Carolina: *CreateSpace Independent Publishing Platform*.

Hawkins, A. J., & Ooms, T. (2012). Can marriage and relationship education be an effective policy tool to help low-income couples form and sustain healthy marriages and relationships? A review of lessons learned. *Marriage & Family Review, 48*, 524–554.

Hayes, R. M., Carter, P. R., Gollop, N. D., Reynolds, J., Uppal, H., Sarma, J., Chandran, S., & Potluri, R. (2016). The impact of marital status on mortality and length of stay in patients admitted with m myocardial infarction. *Heart, 102*(6), A1-A147.

Hill, M. E. (2020). "You can have it all, just not at the same time": Why doctoral students are actively choosing singlehood. *Gender Issues.*

Hill, R. M., Oosterhoff, B., & Do, C. (2019). Using machine learning to identify suicide risk: a classification tree approach to prospectively identify adolescent suicide attempters. *Archives of Suicide Research*, 1-18.

Ho, T. K. (1995, August). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition 1*, 278-282.

Hoffmann, J. P. (2018). Cohabitation, Marijuana Use, and Heavy Alcohol Use in Young Adulthood. *Substance use & misuse*, *53*(14), 2394-2404.

Hohmann-Marriott, B. E. (2006). Shared beliefs and the union stability of married and cohabiting couples. *Journal of Marriage and Family*, *68*(4), 1015-1028.

Hsueh, J., Alderson, D. P., Lundquist, E., Michalopoulos, C., Gubits, D., Fein, D., & Knox, V. (2012). The supporting healthy marriage evaluation: Early impacts on low-income families. *SSRN Electronic Journal*. 10.2139/ssrn.2030319

Isen, A. & Stevenson, B. (2010). Women's education and family behavior: Trends in marriage, divorce and fertility. University of Pennsylvania ScholarlyCommons.

Jamison, T. B. (2018). Cohabitation transitions among low-income parents: A qualitative investigation of economic and relational motications. *Journal of Family Economics, 39*, 73-87

Janson, C., Leisenring, W., Cox, C., Termuhlen, A. M., Mertens, A. C., Whitton, J. A., ... & Kadan-Lottick, N. S. (2009). Predictors of marriage and divorce in adult survivors of childhood cancers: a report from the Childhood Cancer Survivor Study. *Cancer Epidemiology and Prevention Biomarkers*, *18*(10), 2626-2635.

Jemal, A., Ward, E., Anderson, R. N., Murray, T., & Thun, M. J. (2008). Widening of socioeconomic inequalities in US death rates, 1993–2001. *PloS one*, *3*(5).

Johnson, M. K., & Mollborn, S. (2009). Growing up faster, feeling older: Hardship in childhood and adolescence. *Social psychology quarterly*, *72*(1), 39-60.

Kalmijn, M. (1999). Father involvement in childrearing and the perceived stability of marriage. *Journal of Marriage and the Family*, 409-421.

Kefalas, M. J., Furstenberg, F. F., Carr, P. J., Naplitano, L. (2011). Marriage is more than being together: The meaning of marriage for young adults. *Journal of Family Issues, 32*(7), 845-875.

Kendler, H. H. (1987). A good divorce is better than a bad marriage. *Annals of Theoretical Psychology*, *5*, 55-89.

Kerpelman, J. L. (2012). "Relationship Smart" youth: A statewide study of relationship education for high school students. Paper presented at the National Council on Family Relations Annual Conference, October 31, Phoenix, AZ.

Kerr, D., Moyser, M., & Beaujot, R. (2006). Marriage and cohabitation: Demographic and socioeconomic differences in Quebec and Canada. *Canadian Studies in Population [ARCHIVES]*, *33*(1), 83-117.

King, S. M., Iacono, W. G., & McGue, M. (2004). Childhood externalizing and internalizing psychopathology in the prediction of early substance use. *Addiction, 99*(12), 1548-1559.

Knox, J. (2018). *The decline and revival of friendship in America*. Medium. https://medium.com/@johnknox_uab88/the-decline-and-revival-of-friendship-in-america-340bfacc2735

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai 14*(2) 1137-1145.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial intelligence Applications in Computer Engineering, 160*, 3-24.

Kuhn, M. & Quinlan, R. (2020). *C50: C5.0 Decision trees and rule-based models. R package version 0.1.3.1.* https://CRAN.R-project.org/package=C50

Lacey, R. E., Kumari, M., & Bartley, M. (2014). Social isolation in childhood and adult inflammation: Evidence from the National Child Development Study. *Psychoneuroendocrinology*, *50*, 85-94.

Landale, N. S., Schoen, R., & Daniels, K. (2010). Early family formation among White, Black, and Mexican American women. *Journal of Family Issues*, *31*(4), 445-474.

Laplant, K. M. (2016). Singlehood. *Encyclopedia of Family Studies,* 1-5.

Larson, P. J. & Olson, D. H. (2004). Spiritual beliefs and marriage: A national survey based on ENRICH. *The Family Psychologist, 20*(2), 4-8.

Lehrer, E. L. (2000). Religion as a determinant of entry into cohabitation and marriage. In L. J. Waite & C. Bachrach (Eds.), *The ties that bind: Perspectives on Marriage and Cohabitation*, 227-252.

Lichter, D. T. (2001). Marriage as public policy. *Progressive Policy Institute*.

Lichter, D. T., Price, J. P., Swigert, J. M. (2019). Mismatches in the marriage market. *Journal of Marriage and Family, 82, 2*. 796-809.

Lichter, D. T., Turner, R. N., & Sassler, S. (2010). National estimates of the rise in serial cohabitation. *Social Science Research, 39*(5), 754-765.

Lin, I. F., & Wu, H. S. (2019). Sibling Influences, Sibling Similarities, and Parent Care in Late Life. *EurAmerica, 49*(1).

Liu, C. P., Fan, M. Y., Wang, G. W., & Ma, S. L. (2008). Optimizing parameters of support vector machine based on gradient algorithm. *Control and Decision, 23*(11), 1291-1296.

Liverpool, L. (2018). *A bad marriage can seriously damage your health, says, scientist.* The Guardian. https://www.theguardian.com/lifeandstyle/2018/jul/16/a-bad-marriage-is-as-unhealthy-as-smoking-or-drinking-say-scientists

Long, C. R., Seburn, M., Averill, J. R., & More, T. A. (2003). Solitude experiences: Varieties, settings, and individual differences. *Personality and Social Psychology Bulletin*, *29*(5), 578-583. Can increase productivity, experiences of creativity

Lundberg, S., Pollak, R. A., & Stearns, J. (2016). Family inequality: Diverging patterns in marriage, cohabitation, and childrearing. *Journal of Economic Perspectives, 30*(2), 70-102.

Lye, D. N. & Waldron, I. (1998). Relationships of substance use to attitudes toward gender roles, family and cohabitation. *Journal of Substance Abuse*, *10*(2), 185-198.

Manning, E. D. (2020). Young adult relationships in an era of uncertainty: A case for cohabitation. *Demography.*

Manning, W. D. & Smock, P. J. (1995). Why marry? Race and the transition to marriage among cohabitors. *Demography*, *32*(4), 509-520.

Manning, W. D., Smock, P. J., & Fettro, M. N. (2019). Cohabitation and marital expectations among single millennials in the U.S. population research and policy review, *38*(3), 327-364.

Marks, L. (2008). How does religion influence marriage? Christian, Jewish, Mormon and Muslim perspectives. *Journal of Marriage & Family Review, 38*(1), 85-111.

Martino, S. C., Collins, R. L., & Ellickson, P. L. (2004). Substance use and early marriage. *Journal of Marriage and Family*, *66*(1), 244-257.

Mastekaasa, A. (2006). Is marriage/cohabitation beneficial for young people? Some evidence on psychological distress among Norwegian college students. *Journal of Community & Applied Social Psychology, 16*(2), 149-165.

Menasco, M. A. & Blair, S. L. (2014). Adolescent substance use and marital status in adulthood. *Journal of Divorce & Remarriage*, *55*(3), 216-238.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). *E1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. R package version 1.7-3*. https://CRAN.R-project.org/package=e1071

Modell, J. (1980). Normative aspects of American marriage timing since World War II. *Journal of Family History, 5*(2), 210-234.

Morris, W. L., Sinclair, S., & DePaulo, B. M. (2007). No shelter for singles: The perceived legitimacy of marital status discrimination. *Group Processes & Intergroup Relations, 10*(4), 457-470.

Murray, M. (2016). Obergefell v. Hodges and Nonmarriage Inequality. *California Law Review*, 1207-1258.

Musick, K., Brand, J. E., & Davis, D. (2012). Variation in the relationship between education and marriage: Marriage market mismatch?. *Journal of Marriage and Family*, *74*(1), 53-69.

Musick, K., & Bumpass, L. (2012). Reexamining the case for marriage: Union formation and changes in well-being. *Journal of Marriage and Family*, *74*(1), 1-18.

Newport, F. (2009). *Marriage remains key predictor of party identification: Married Americans tilt republican; Unmarried Americans, democratic.* Gallup. https://news.gallup.com/poll/121571/marriage-remains-key-predictor-party-identification.aspx

Nock, S. L. (2005). Marriage as a public issue. *The Future of Children*, 13-32.

Owen, J., Rhoades, G. K., & Stanley, S. M. (2013). Sliding versus deciding in relationships: Associations with relationship quality, commitment, and infidelity. *Journal of Couple & Relationship Therapy*, *12*(2), 135-149.

Peplau, L. A., Hill, C. T., & Rubin, Z. (1993). Sex role attitudes in dating and marriage: A 15-year follow-up of the Boston Couples Study. *Journal of Social Issues, 49*(3), 31-52.

Perelli-Harris, B., Styrc, M., Addo, F., Hoherz, S., Lappegard, T., Sassler, S., & Evans, A. (2017). Comparing the benefits of cohabitation and marriage for health in mid-life: Is the relationship similar across countries? *Economic & Social Research Council: Centre for Population Change*. ISSN: 2042-4116

Pollar, K. M., Jacobsen, L. A., & Mather, M. (2020). The U.S. population is growing at the slowest rate since the 1930s. https://www.prb.org/the-u-s-population-is-growing-at-the-slowest-rate-since-the-1930s/

Pudrovska, T., Schieman, S., & Carr, D. (2006). Strains of singlehood in later life: Do race and gender matter?. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *61*(6), S315-S322.

Quinlan, J. R. (1993). Induction of decision trees. In *Readings in knowledge acquisition and learning: Automating the construction and improvement of expert systems* (pp. 349-361).

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1(*1), 81-106.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rackin, H. M. & Gibson-Davis, C. M. (2017). Low-income childless young adults' marriage and fertility frameworks. *Journal of Marriage and Family, 79*(4), 1096-1110.

Rebecca, L. (1996). Childhood sexual abuse and adult loneliness and network orientation. *Child Abuse & Neglect*, *20*(11), 1087-1093.

Ressler, R. W. & Waters, M. S. (1995). The economics of cohabitation. *Kyklos, 48*(4), 577-592.

Rettner, R. (2012). Marriage, cohabitation provide similar health benefits. *Live Science: Health*, Published January 19, 2012.

Rhoades, G. K., Stanley, S. M., Markman, H. J. (2009). Couples' reasons for cohabitation: associations with individual well-being and relationship quality. *Journal of Family Issues, 30(*2), 233-258.

Rhoades, G. K., Stanley, S. M., Markman, H. J. (2012). A longitudinal investigation of commitment dynamics in cohabiting relationships. *Journal of Family Issues, 33*(3), 369-390.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, 172-181.

Romero, A. P. (2017). *1.1 Million LGBT adults are married to someone of the same sex at the two-year anniversary of Obergefell v. Hodges*. Williams Institute, UCLA School of Law.

RuleQuest. (2019). Data mining tool See5 and C5.0. https://www.rulequest.com/see5-info.html

Sarkisian, N. & Gerstel, N. (2016). Does singlehood isolate or integrate? Examining the link between marital status and ties to kin, friends, and neighbors. *Journal of Social and Personal Relationships*, *33*(3), 361-384.

Schneider, D. (2011). Wealth and the marital divide. *American Journal of Sociology*, *117*(2), 627-667.

Schultz, T. W. (1974). *Economics of the family; marriage, children and human capital; a conference report*. University of Chicago Press, Chicago, US.

Schwartz, J. (2005).The socio-economic benefits of marriage: a review of recent evidence from the United States. *Economic Affairs, 25*(3), 45-51.

Segrin, C., McNelis, M., & Pavlich, C. A. (2018). Indirect effects of loneliness on substance use through stress. *Health Communication*, *33*(5), 513-518.

Segrin, C., Nevarez, N., Arroyo, A., & Harwood, J. (2012). Family of origin environment and adolescent bullying predict young adult loneliness. *The Journal of Psychology*, *146*(1-2), 119-134.

Seltzer, J. A. (2000). Families formed outside of marriage. *Journal of Marriage and Family, 62*(4), 1247-1268.

Shafer, K., & James, S. L. (2013). Gender and socioeconomic status differences in first and second marriage formation. *Journal of Marriage and Family*, *75*(3), 544-564.

Shmerling, R. H. (2016). *The health advantages of marriage*. Harvard Health Publishing: Harvard Medical School. https://www.health.harvard.edu/blog/the-health-advantages-of-marriage-2016113010667

Silva, D., Goulart, P., & Obreshkova, E. (2016). Marriage rates. *Encyclopedia of Family Studies,* 1-6.

Smock, P. J. & Schwartz, C. R. (2020). The demography of families: A review of patterns and change. *Journal of Marriage and Family*, *82*(1), 9-34.

Soons, J. P., Liefbroer, A. C., & Kalmijn, M. (2009). The long-term consequences of relationship formation for subjective well-being. *Journal of Marriage and Family, 71*(5), 1254-1270.

Stanley, S. M. (2001). Making a case for premarital education. *Family Relations, 50*, 272–280.

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. B*MC Bioinformatics, 9*(1), 319.

Steinberg-Schone, B. (1998). Health-related behaviors and the benefits of marriage for elderly persons. *The Gerontologist, 35*(5), 618-627.

Stepler, R. (2017). *Number of US adults cohabitating with a partner continues to rise, especially among those 50 and older*. Pew Research Center. https://www.pewresearch.org/fact-tank/2017/04/06/number-of-u-s-adults-cohabiting-with-a-partner-continues-to-rise-especially-among-those-50-and-older/

Stern, B. (2019). Mass Incarceration: The Subjugation of Black Men's Health.

Stevens, G., Owens, D., & Schaefer, E. C. (1990). Education and attractiveness in marriage choices. *Social Psychology Quarterly*, 62-70.

Stutzer, A., & Frey, B. S. (2006). Does marriage make people happy, or do happy people get married?. *The Journal of Socio-Economics*, *35*(2), 326-347.

Suthaharan, S. (2016). Machine learning models and algorithms for big data classification. Integr. *Ser. Inf. Syst, 36,* 1-12.

Swartz, T. T., Kim, M., Uno, M., Mortimer, J., & O'Brien, K. B. (2011). Safety nets and scaffolds: Parental support in the transition to adulthood. *Journal of Marriage and Family*, *73*(2), 414-429.

Tang, C., Curran, M., & Arroyo, A. (2014). Cohabitors' reasons for living together, satisfaction with sacrifices, and relationship quality. *Journal of Marriage and Family Review, 50*(7), 598-620.

Teachman, J. (2003). Premarital sex, premarital cohabitation, and the risk of subsequent marital dissolution among women. *Journal of Marriage and Family*, *65*(2), 444-455.

Teachman, J. D. & Polonko, K. A. (1990). Cohabitation and marital stability in the United States. *Social Forces, 69*(1), 207-220.

The Daily. (2017). Families, households, and marital status: Key results from the 2016 Census. https://www150.statcan.gc.ca/n1/en/daily-quotidien/170802/dq170802a-eng.pdf?st=D1OXhB_X

Therneau, T. (2017). *A package for survival analysis in S. version 2.38. 2015*.

Thomas, A. & Sawhill, I. (2002). For richer or for poorer: Marriage as an anti-poverty strategy. *Journal of Policy Analysis and Management, 21*(4), 587-599.

Thornton, A. (1991). Influence of the marital history of parents on the marital and cohabitational experiences of children. *American Journal of Sociology*, *96*(4), 868-894.

Thornton, A., Axinn, W. G., & Hill, D. H. (1992). Reciprocal effects of religiosity, cohabitation, and marriage. *American Journal of Sociology*, *98*(3), 628-651.

Timberlake, J. M. & Heuveline, P. (2005). Changes in nonmarital cohabitation and the family structure experiences of children across 17 countries. Sociological studies of children and youth, *10*, 257-278.

Tumin, D. (2016). Marriage trends among Americans with childhood-onset disabilities. *Disability and Health Journal, 9*(4), 713-718.

Underwood, K. L. (2013). *Attachment style, relationship satisfaction, and loneliness in African American adults in relation to childhood parental divorce* (Doctoral dissertation, Walden University).

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Vespa, J. & Painter, M. A. (2011). Cohabitation history, marriage, and wealth accumulation. *Demography*, *48*(3), 983-1004.

Waldron, I., Hughes, M. E., & Brooks, T. L. (1996). Marriage protection and marriage selection—prospective evidence for reciprocal effects of marital status and health. *Social Science & Medicine*, *43*(1), 113-123.

Wang, W. & Parker, K. (2014). Record share of Americans have never married: As values, economics, and gender patterns change. Pew Research Center. https://www.pewsocialtrends.org/2014/09/24/record-share-of-americans-have-never-married/

Western, B. & Lopoo, L. (2004). Incarceration, marriage, and family life. *Princeton University, Department of Sociology (paper)*

Whisman, M. A. (2006). Childhood trauma and marital outcomes in adulthood. *Personal Relationships*, *13*(4), 375-386.

Willis, S. L., & Baltes, P. B. (1980). Intelligence in adulthood and aging: Contemporary issues. In L. W. Poon (Ed.), Aging in the 1980s: Psychological issues (p. 260–272). American Psychological Association. https://doi.org/10.1037/10050-019

Willoughby, B. J. & Carroll, J. S. (2010). Sexual experience and couple formation attitudes among emerging adults. *Journal of Adult Development*, *17*(1), 1-11.

Wright, M. N., & Ziegler, A. (2015). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

Wu, Z., & Hart, R. (2002). The effects of marital and nonmarital union transition on health. *Journal of Marriage and Family, 64*(2), 420-432.

Wu, Z., Penning, M. J., Pollard, M. S., & Hart, R. (2003). "In sickness and in health" Does cohabitation count?. *Journal of Family Issues, 24*(6), 811-838.

Xie, Y., Raymo, J. M., Goyette, K., & Thornton, A. (2003). Economic potential and entry into marriage and cohabitation. *Demography*, *40*(2), 351-367.

Yamokoski, A. (2007). Wealth inequality: effects of gender, marital status, and parenthood on asset accumulation (Doctoral dissertation, The Ohio State University).

Yoshida, A. (2011). No chance for romance: Corporate culture, gendered work, and increased singlehood in Japan. *Contemporary Japan*, *23*(2), 213-234.

Zimmerman, A. C. & Easterlin, R. A. (2006). Happily ever after? Cohabitation, marriage, divorce, and happiness in Germany. *Population and Development Review, 32*(3), 511–528.

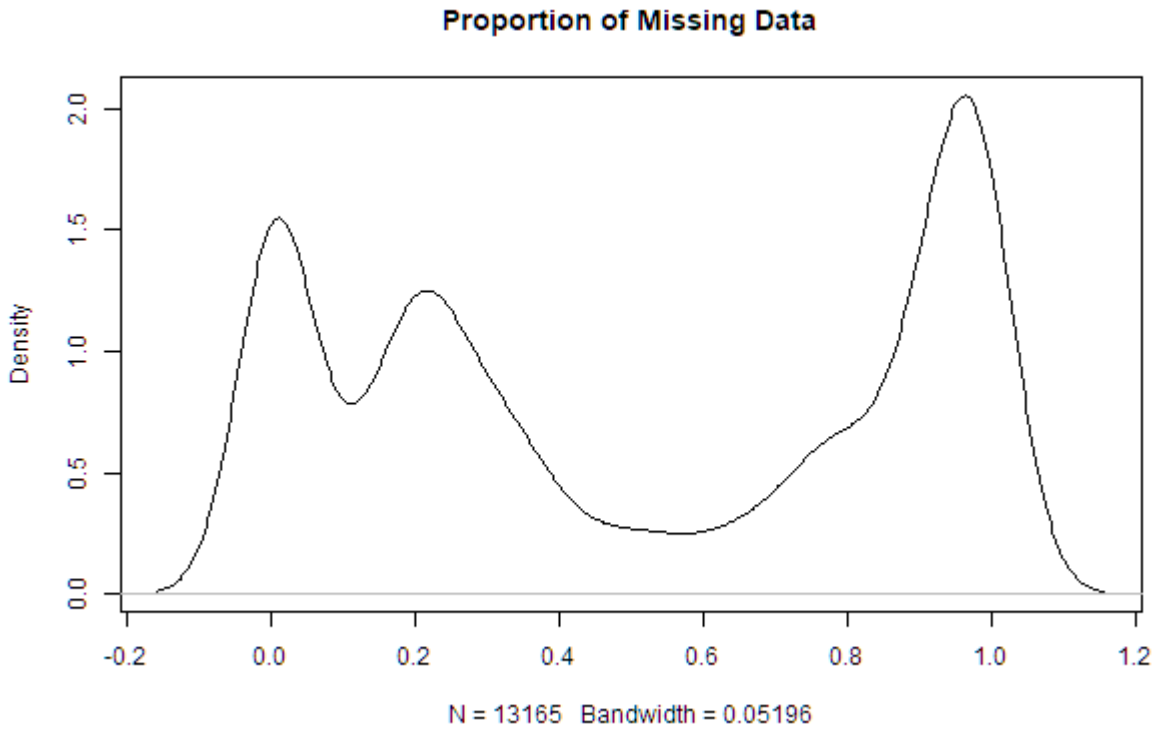
Appendix A

Table A1
Code books removed from the machine learning analysis.

Code Book	Wave
H1HR	Household Roster and Residence History
H2HR	Household Roster and Residence History
H3HR	Household Roster
H4HR	Household Roster
H1RI	Relationship Information-Audio CASI
H2RI	Relationship Information-Audio CASI
H1RX	Non-Relationship History Audio CASI
H2RX	Non-Relationship History Audio CASI
H3TR	Relationships
H4TR	Relationships
H3RD	Relationships in Detail
H4RD	Relationships in Detail
H3MR	Marriage/Co-habitation History and Attitudes
H3LB	Live Births
H3PC	Current Pregnancies
H3PG	Completed Pregnancies
H3KK	Children and Parenting

Note. In addition to removing these code books in their entirety, a number of additional items that referenced marriage, cohabitation, or singlehood were removed from the analysis. For example, H4HS1 was removed because it includes an item asking if "you are covered by your husband's or wife's insurance." In total, fewer than 50 such additional items were flagged for exclusion by hand after a comprehensive search of the Add Health code books.

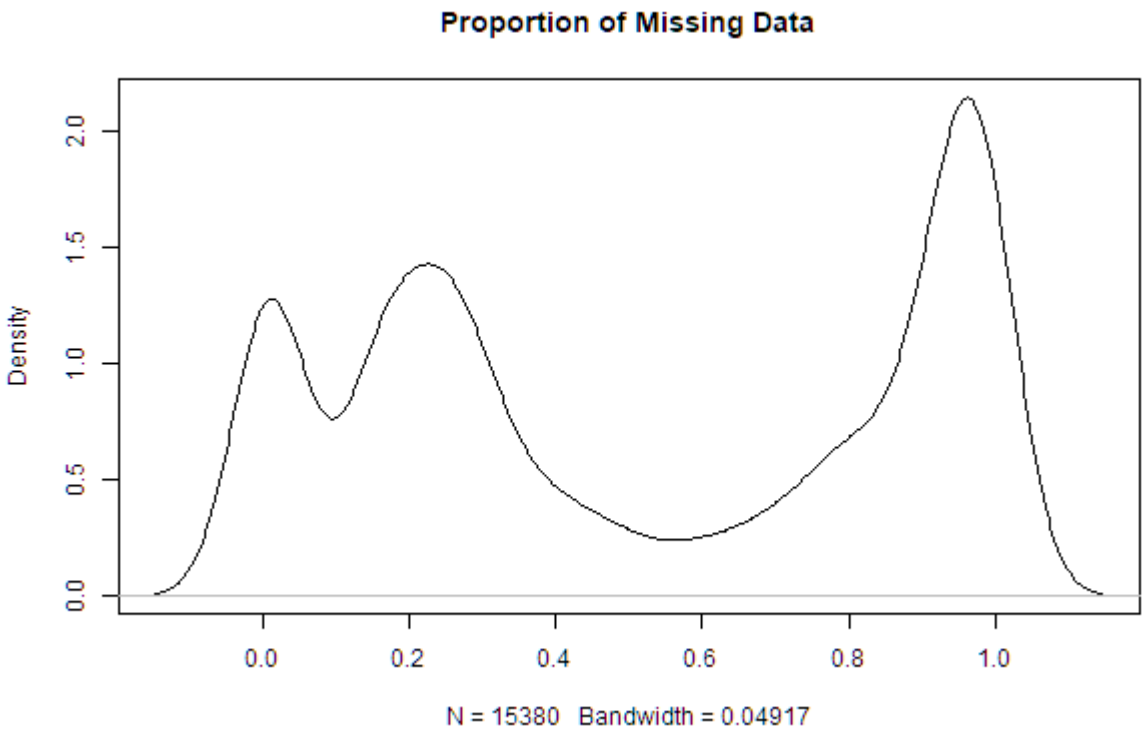
Figure A1
Proportion of Missing Data for Analysis I



Note. This figure shows the proportion of missing data across the semi-preprocessed training variables in Analysis I. Variables with greater than 80% missing (including the peak on the right) were removed from the analysis.

Figure A2

Proportion of Missing Data for Analysis II



Note. This figure shows the proportion of missing data across the semi-preprocessed training variables in Analysis II. Variables with greater than 80% missing (including the peak on the right) were removed from the analysis.

Appendix B

Figure B1

Training accuracy and kappa scores in the training data 5-fold cross-validation for Analysis I.

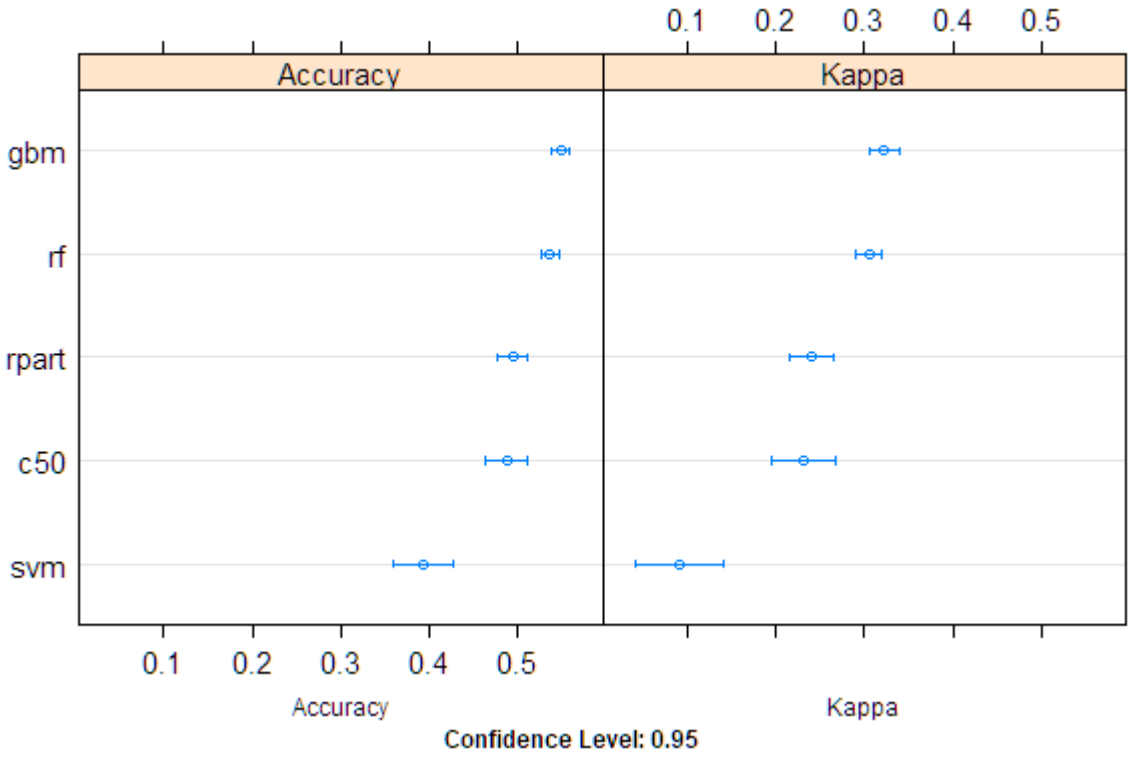


Figure B2

Cross-validated accuracy in the training data for the gbm model using the evaluated hyperparameters in Analysis I.

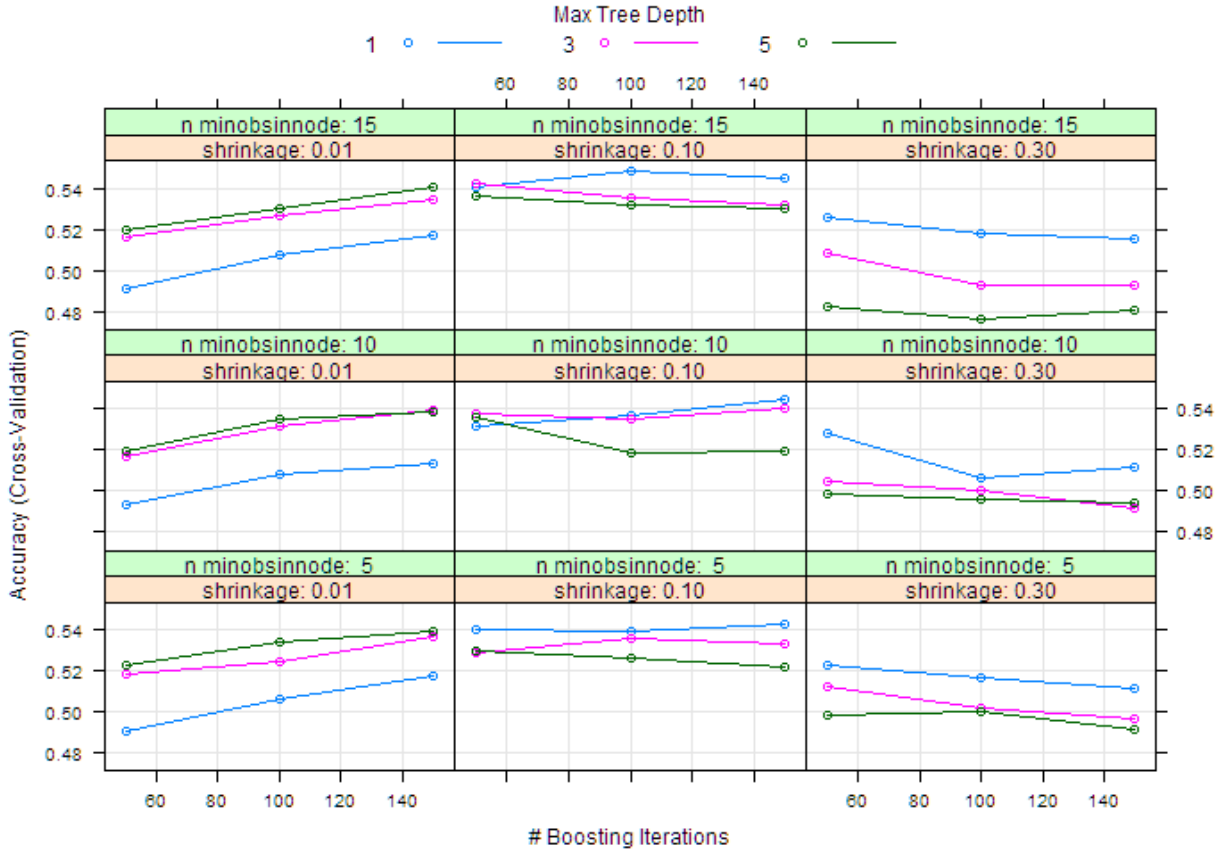


Figure B3

Training accuracy and kappa scores in the training data 5-fold cross-validation for Analysis II.

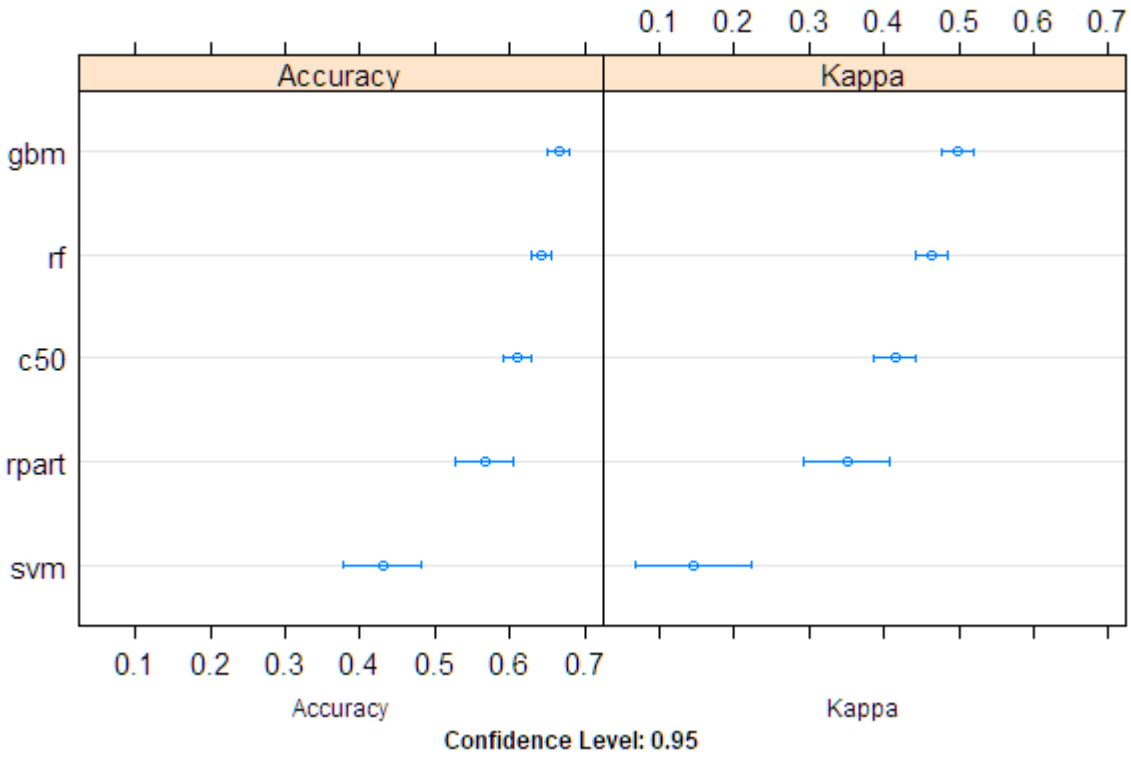
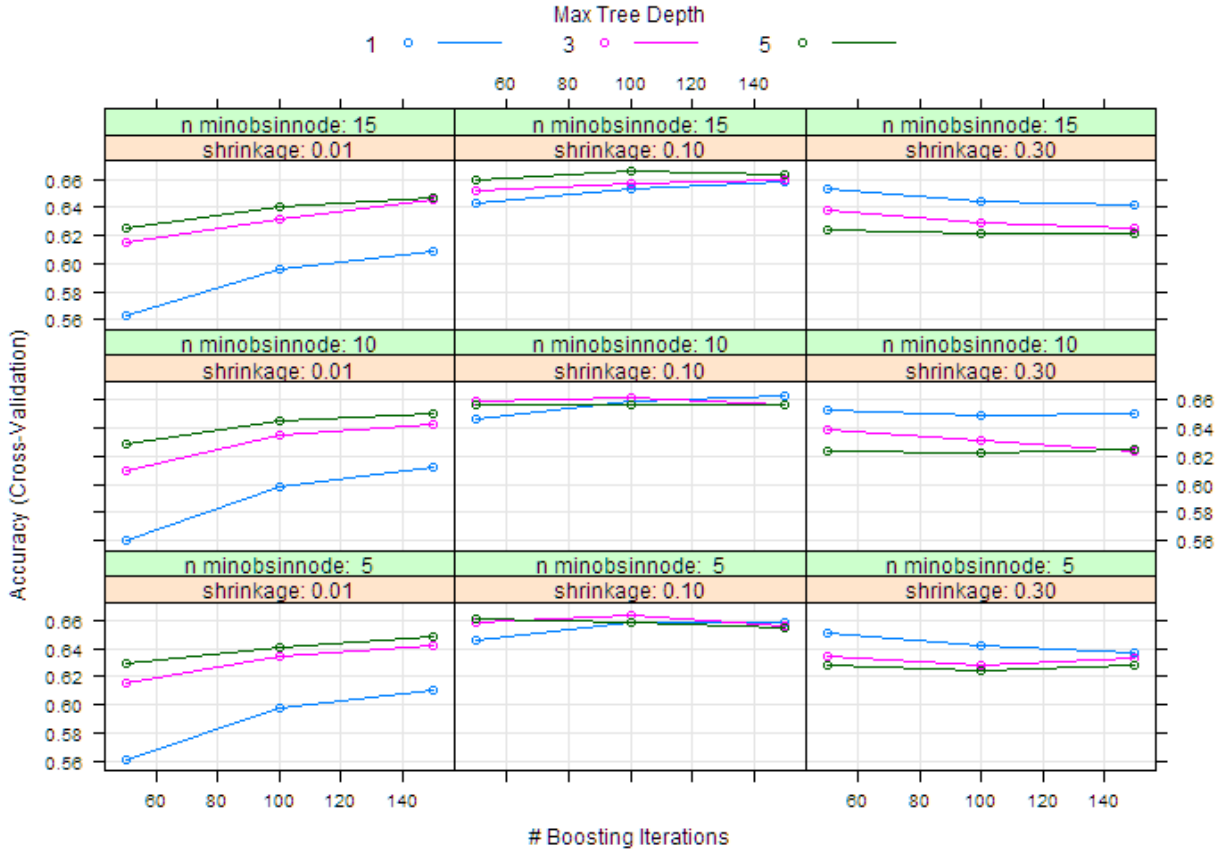
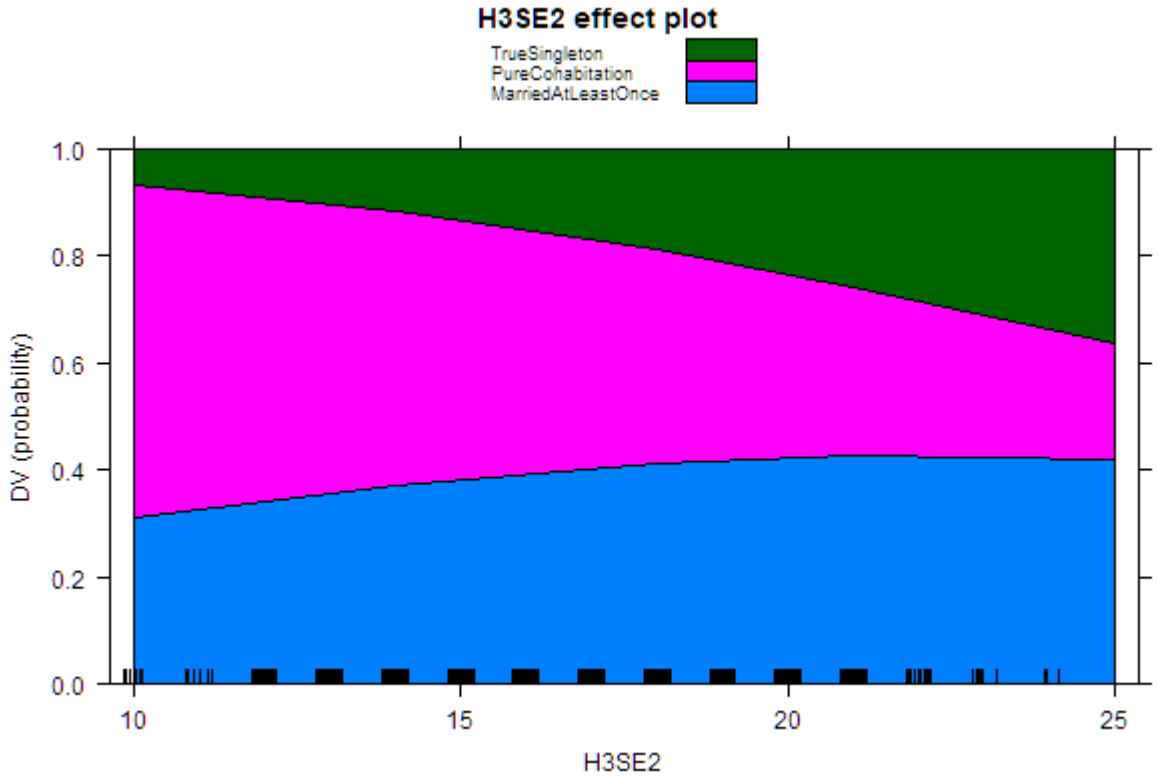


Figure B2

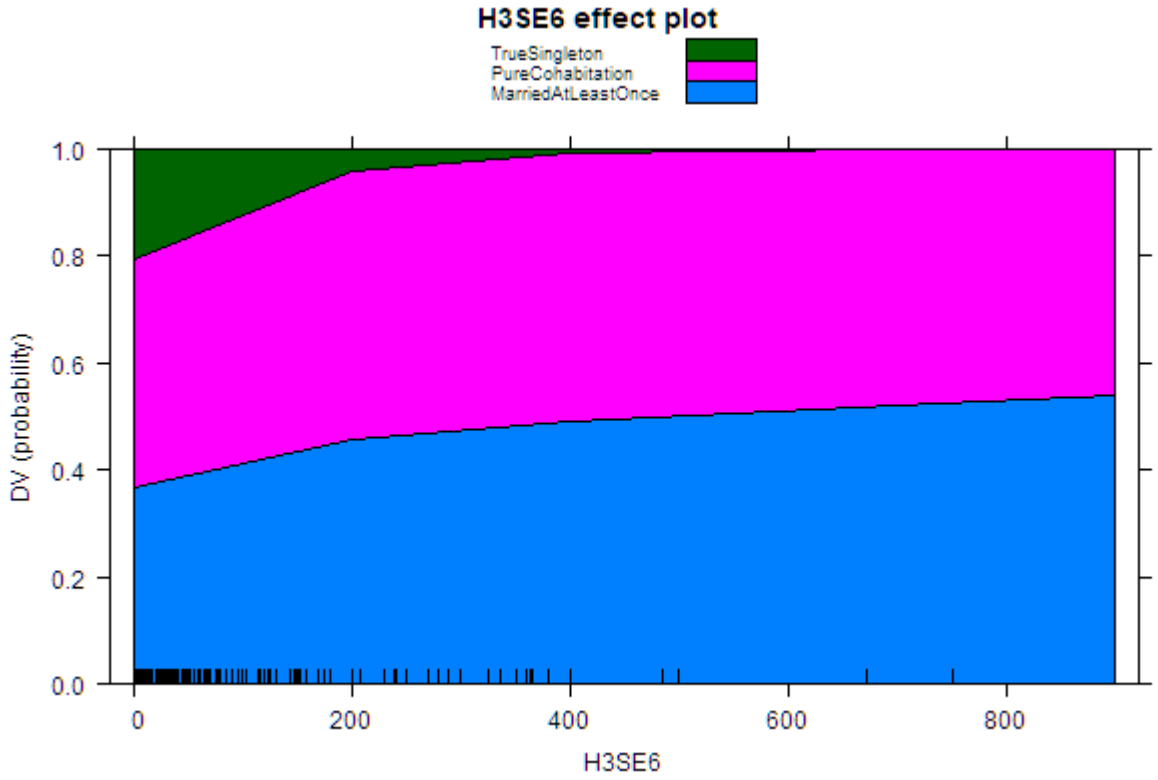
Cross-validated accuracy in the training data for the gbm model using the evaluated hyperparameters in Analysis II.



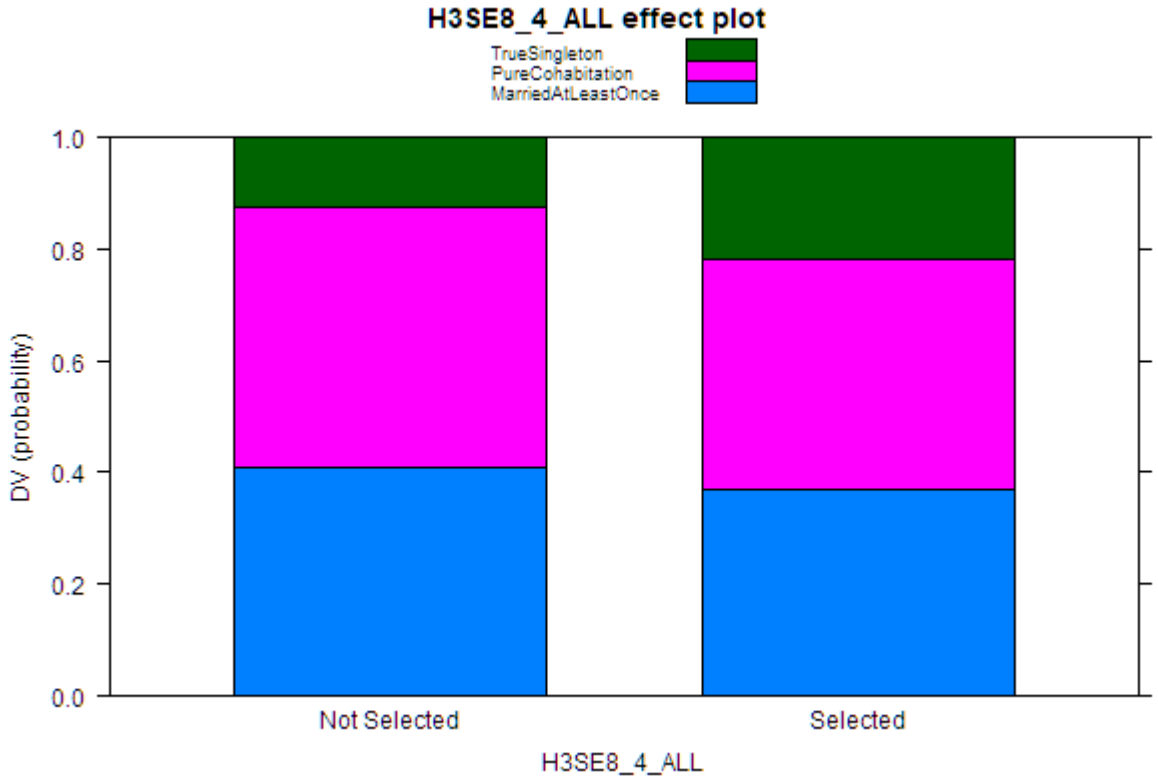
Appendix C



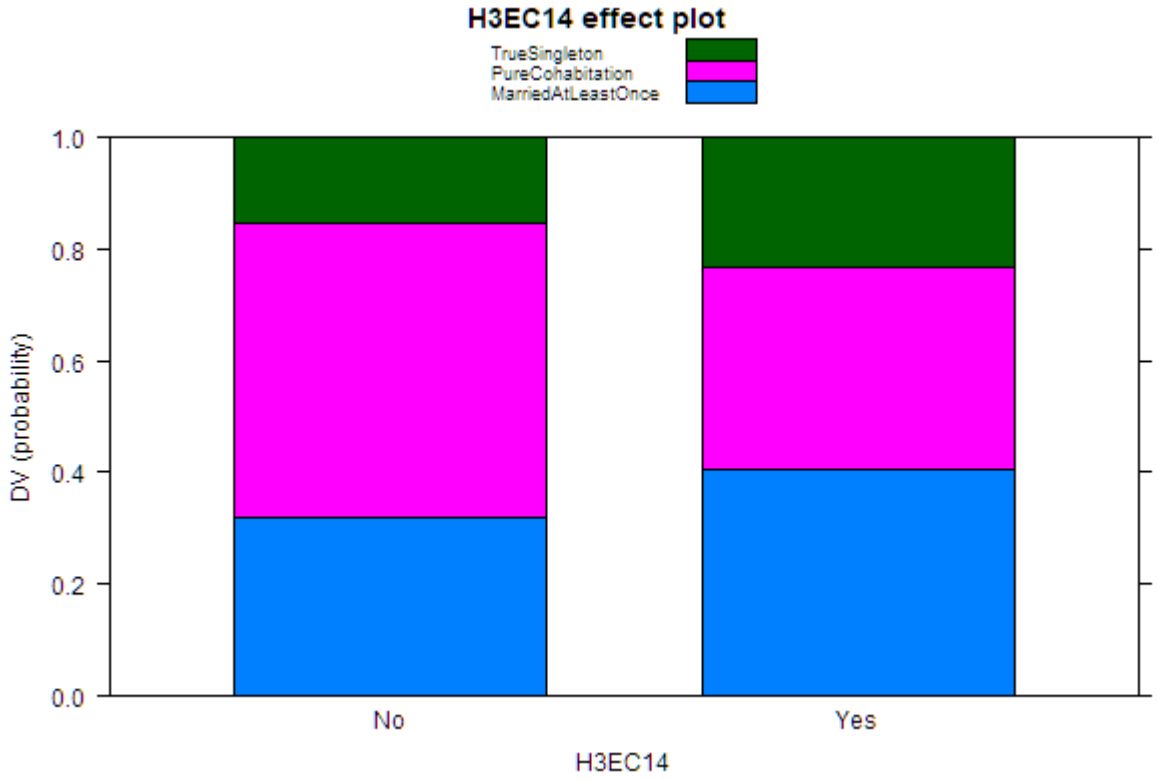
Note. H3SE2 (Wave III): "How old were you the first time you had vaginal intercourse?"



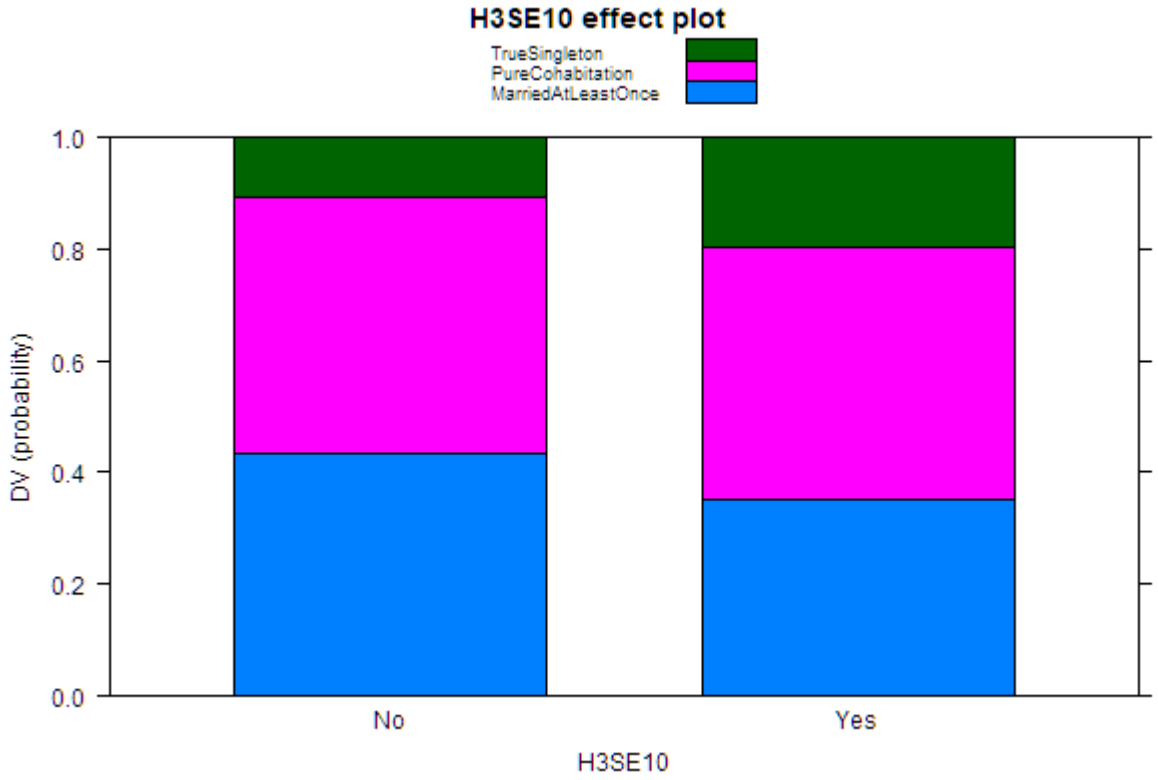
Note. H3SE6 (Wave III): "How many times have you had vaginal intercourse in the past 12 months?"



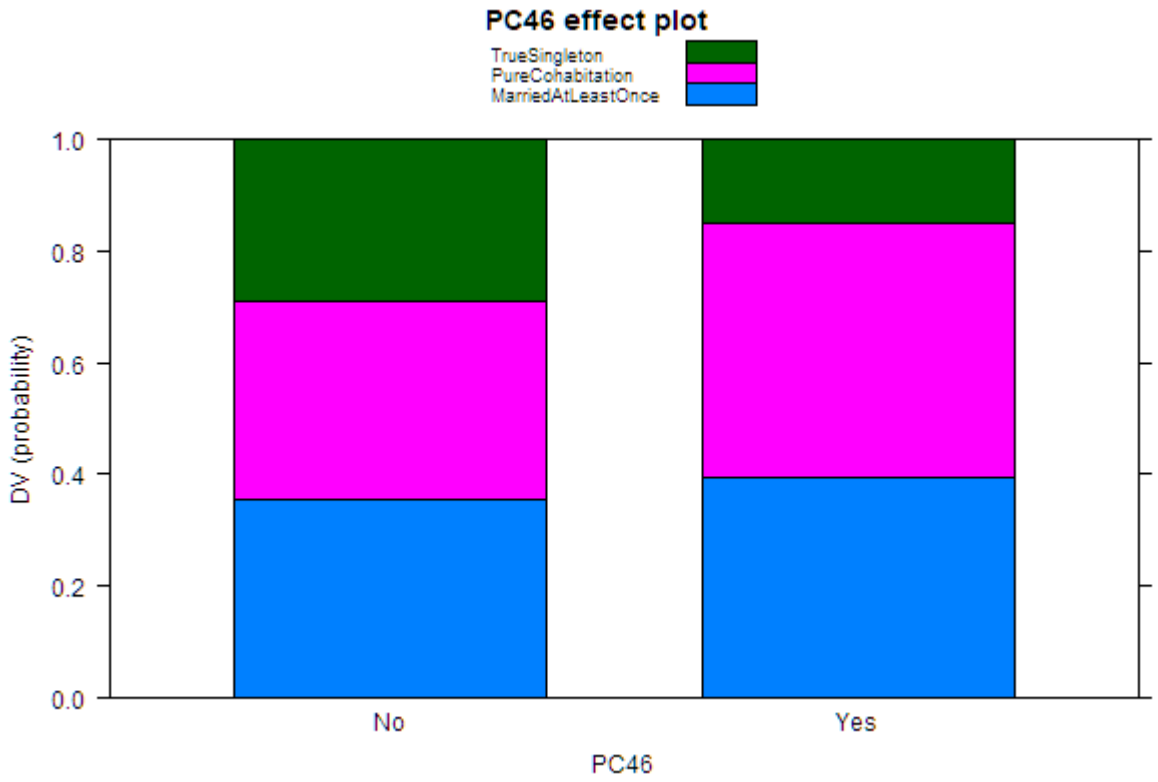
Note. H3SE8 (Wave III): "On how many of these occasions did {YOU/YOUR PARTNER} use a condom?" Response: 4 - all



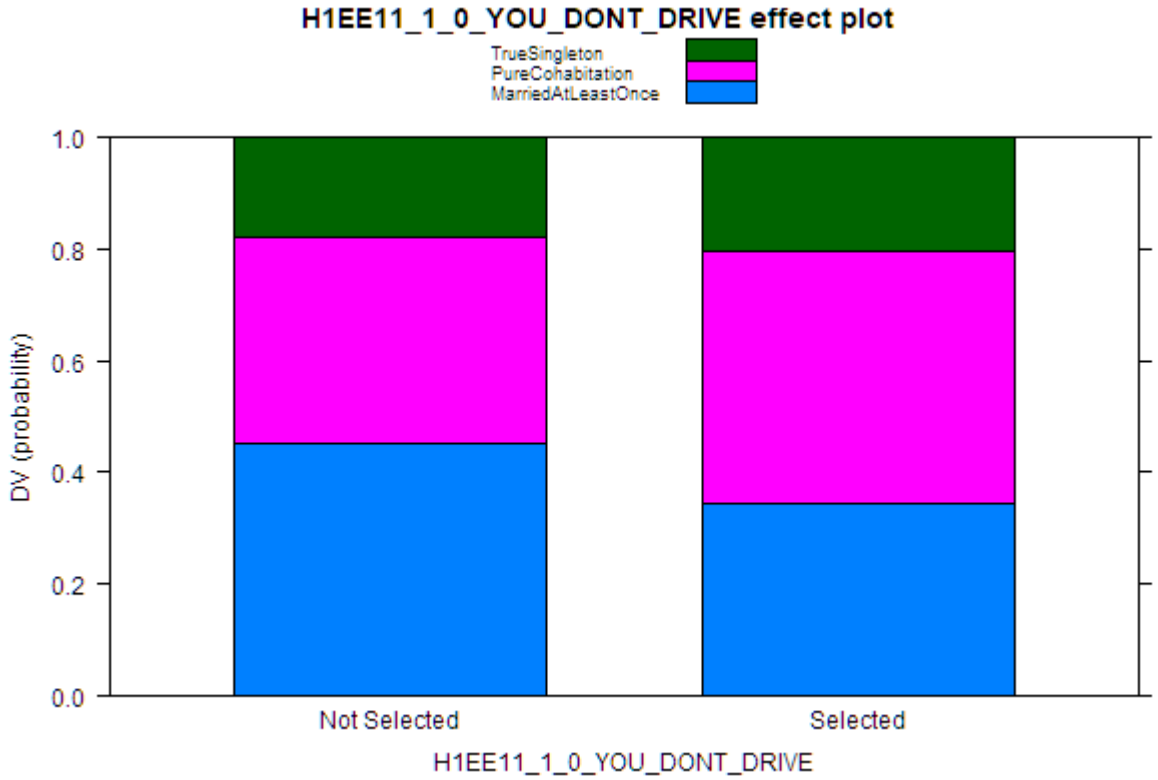
Note. H3EC14 (Wave III): "Do you have an email account?"



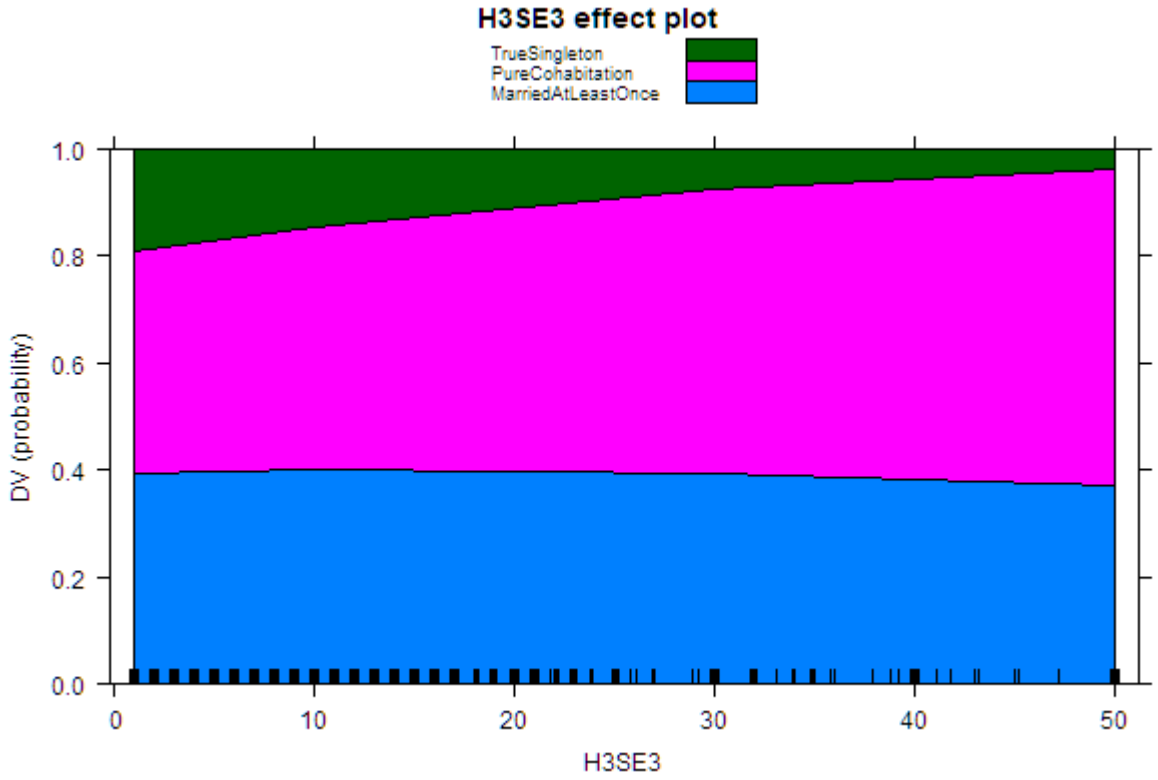
Note. H3SE10 (Wave III): "The most recent time you had vaginal intercourse did {YOU/YOUR PARTNER} use a condom?"



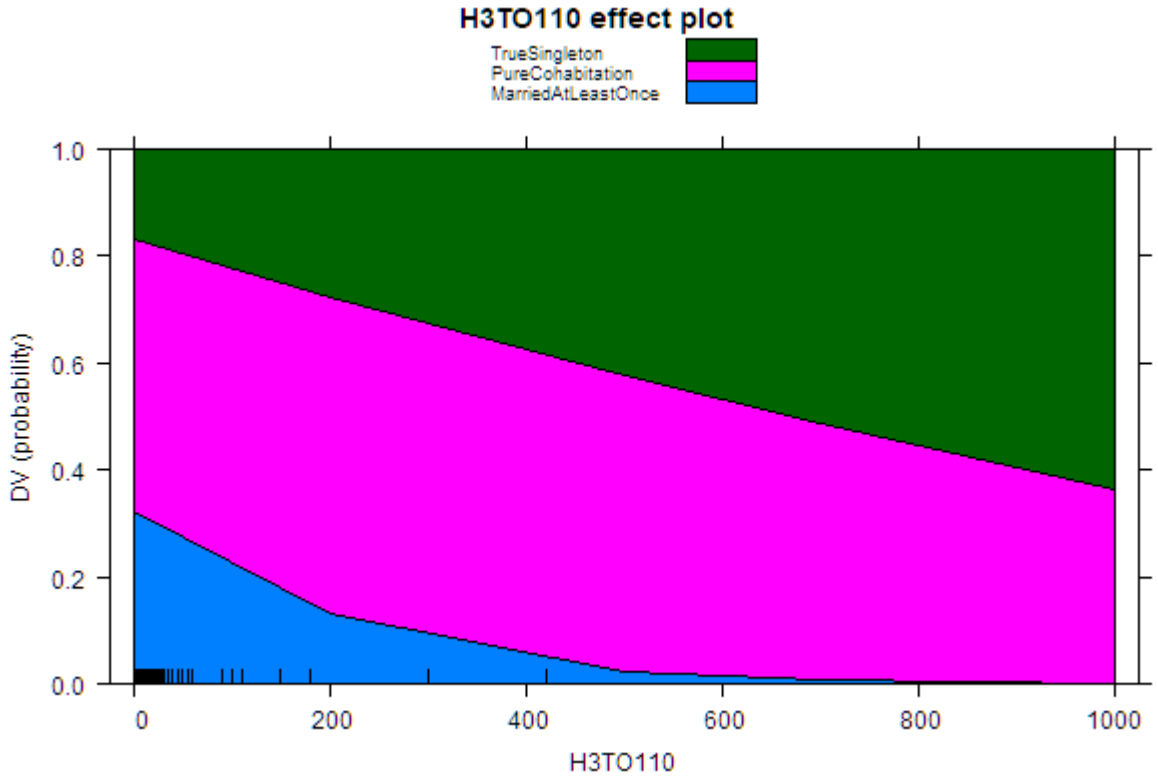
Note. PC46 (Wave I): "Do you think that (he/she) has ever kissed and necked?"



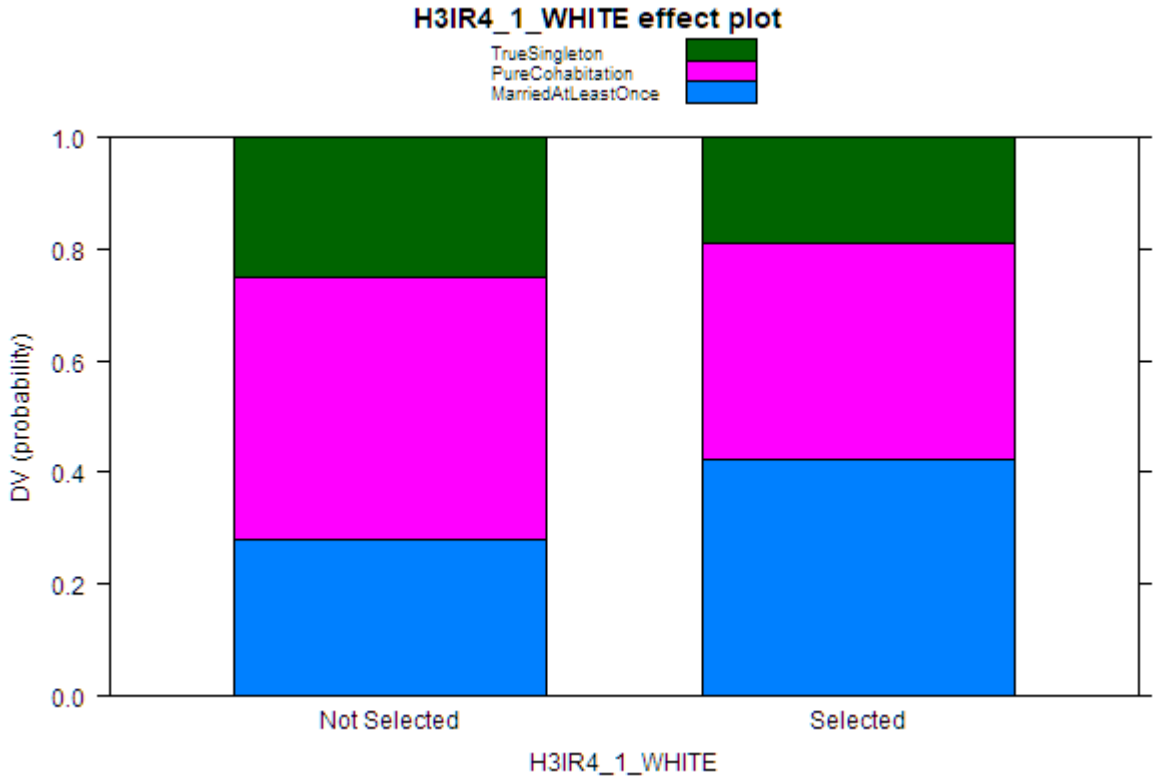
Note. H1EE11 (Wave I): "About how many miles do you drive each week?" Response: 1 - 0, you don't drive



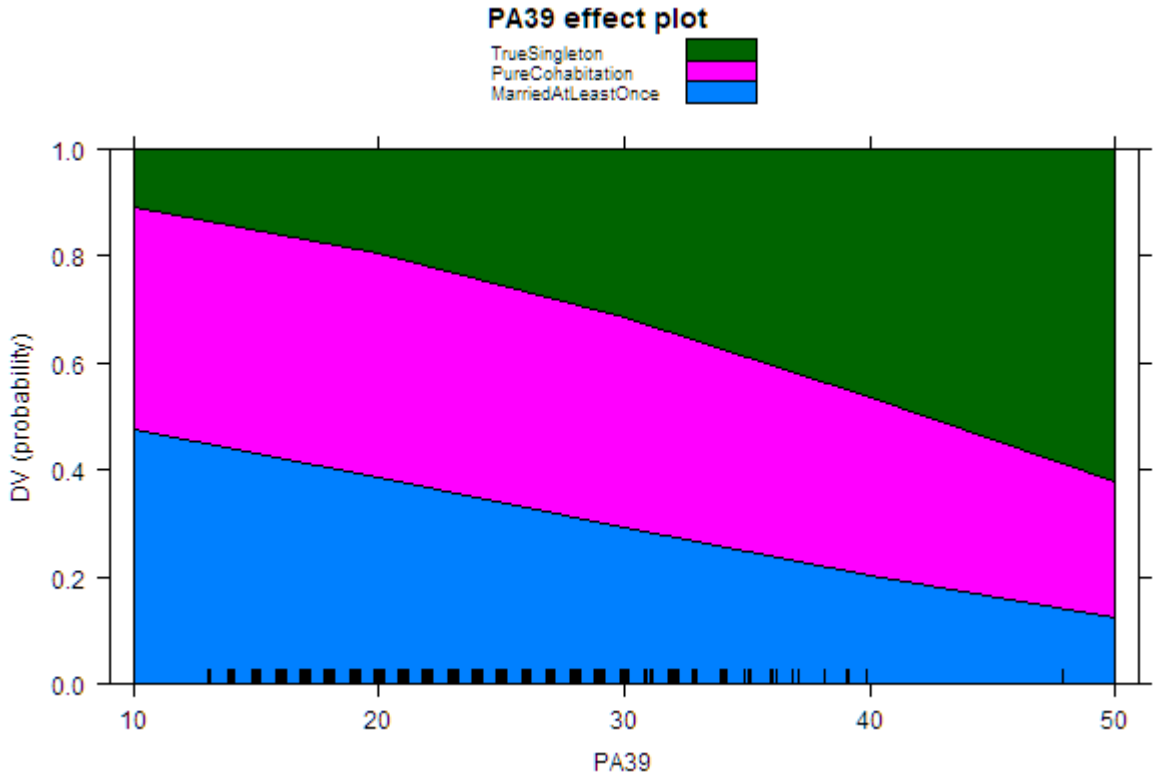
Note. H3SE3 (Wave III): "With how many partners have you ever had vaginal intercourse, even if only once?"



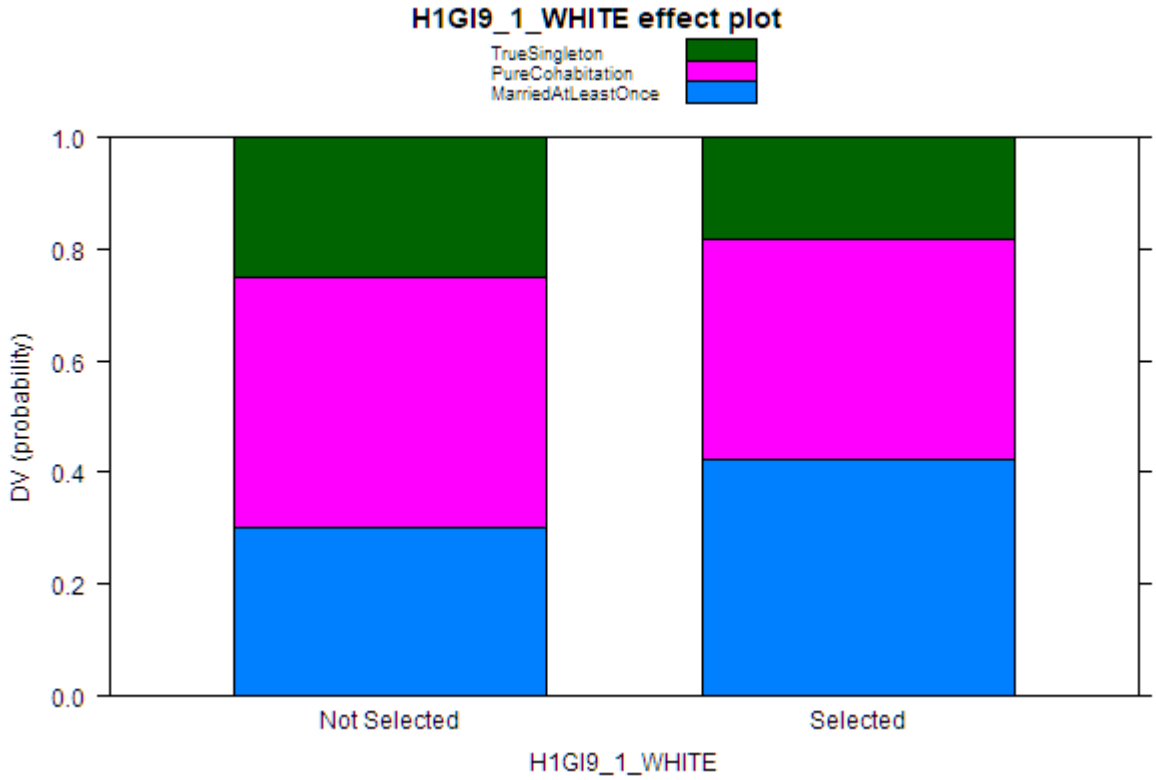
Note. H3TO110 (Wave III): "During the past 30 days, how many times have you used marijuana?"



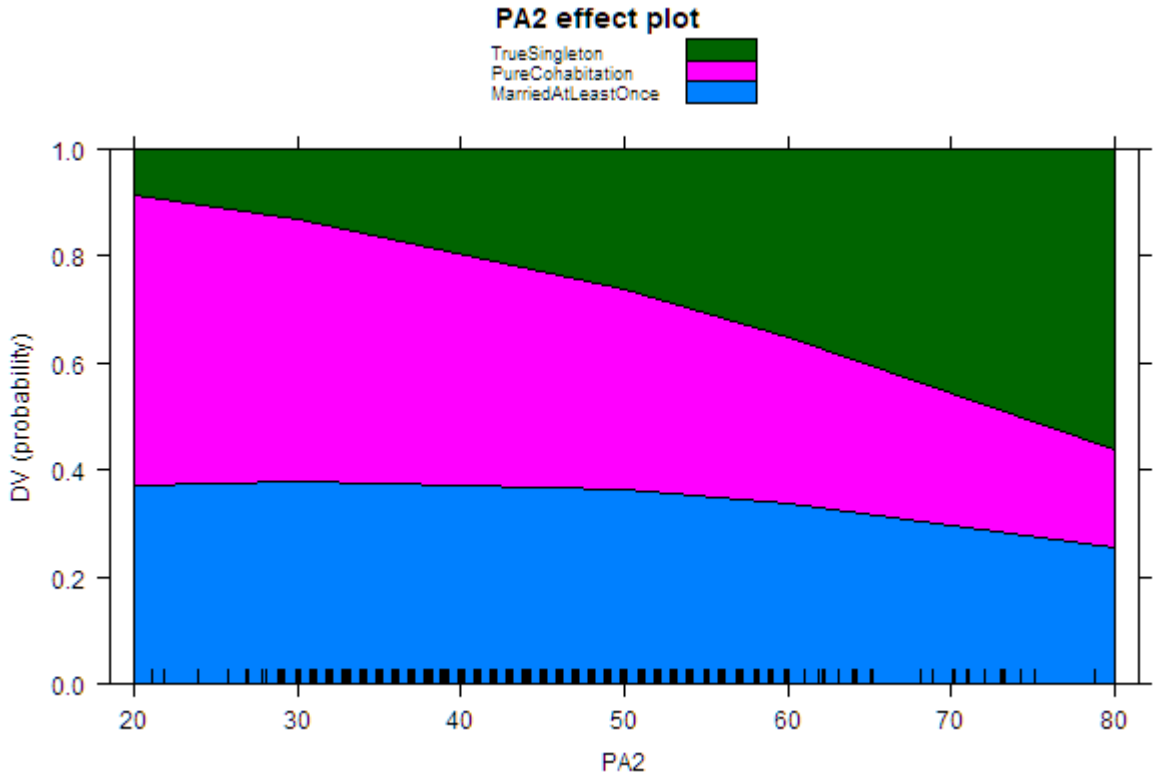
Note. H3IR4 (Wave III): "Indicate the race of the respondent from your own observation (not from what the respondent said)." Response: 1 - White



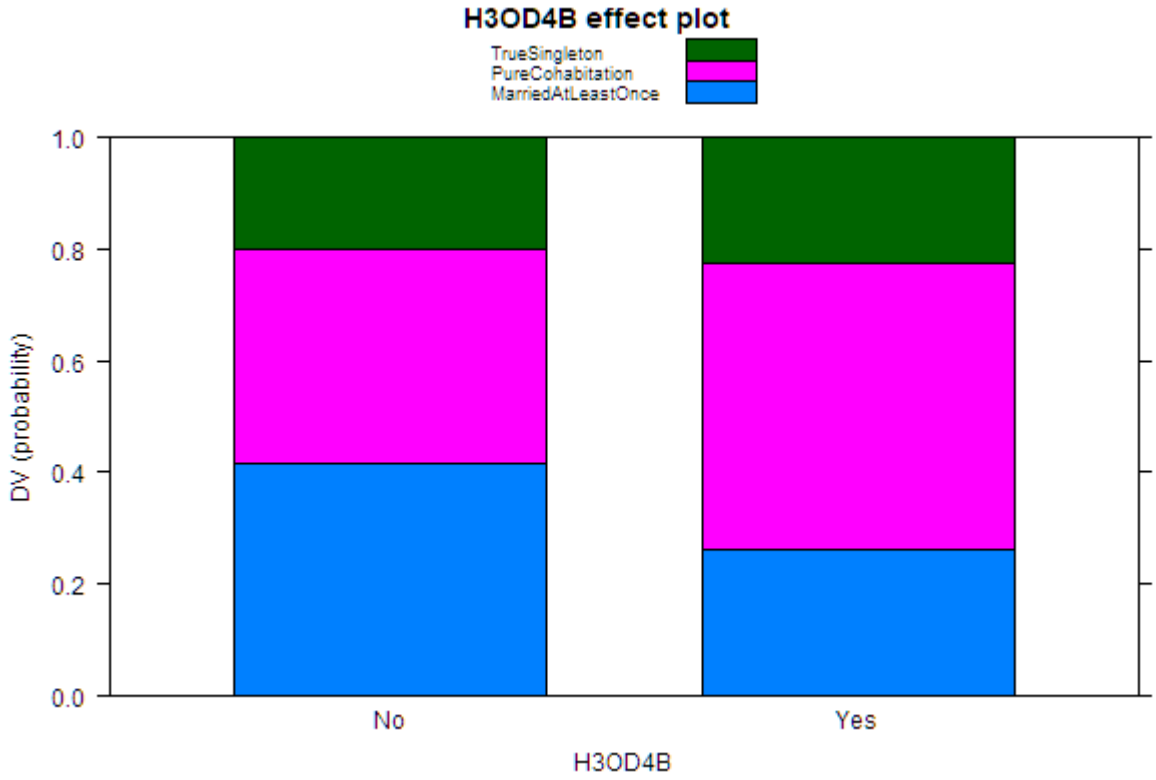
Note. PA39 (Wave I): "How old were you when you were first married?"



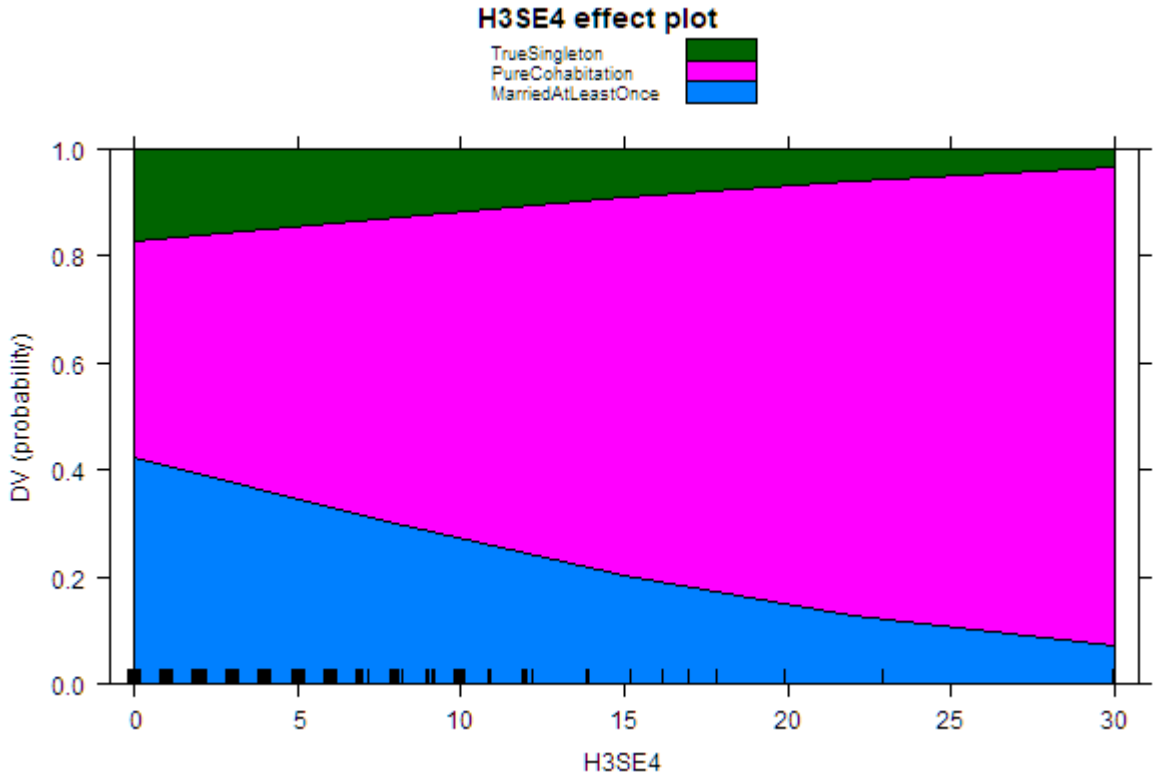
Note. H1GI9 (Wave I): "Interviewer: Please code the race of the respondent from your observation alone." Response: 1 - White



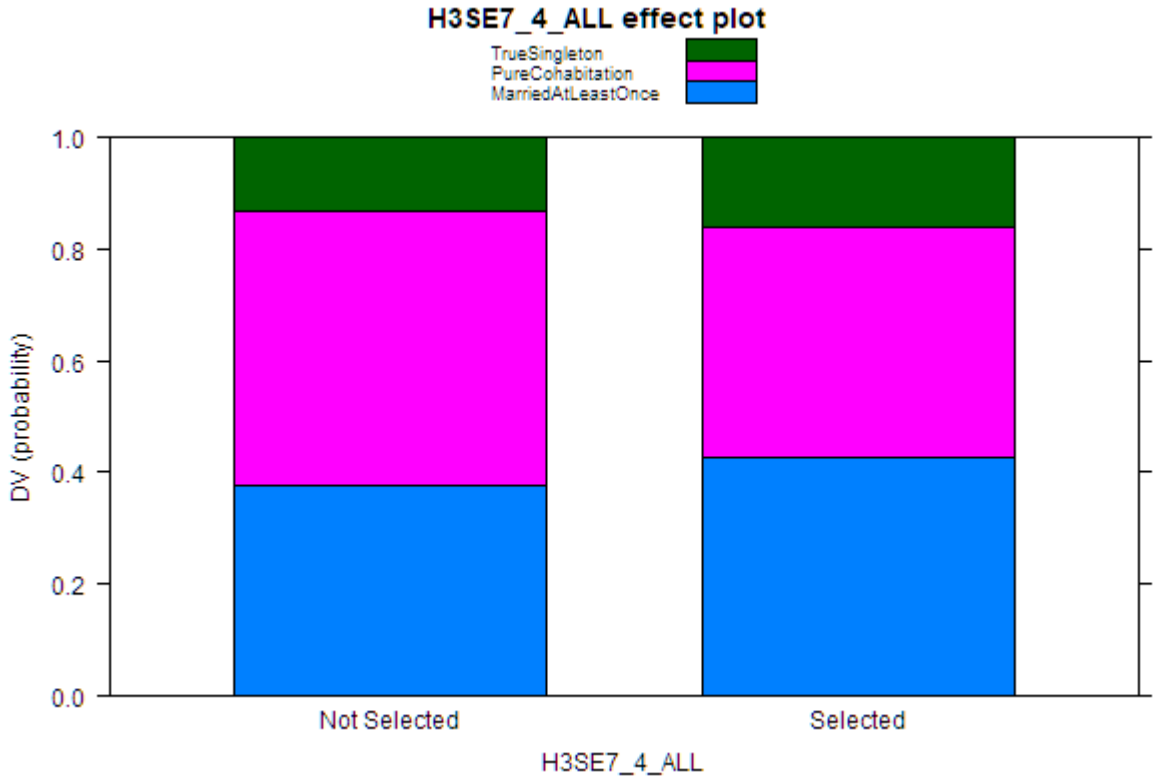
Note. PA2 (Wave I): "How old are you?"



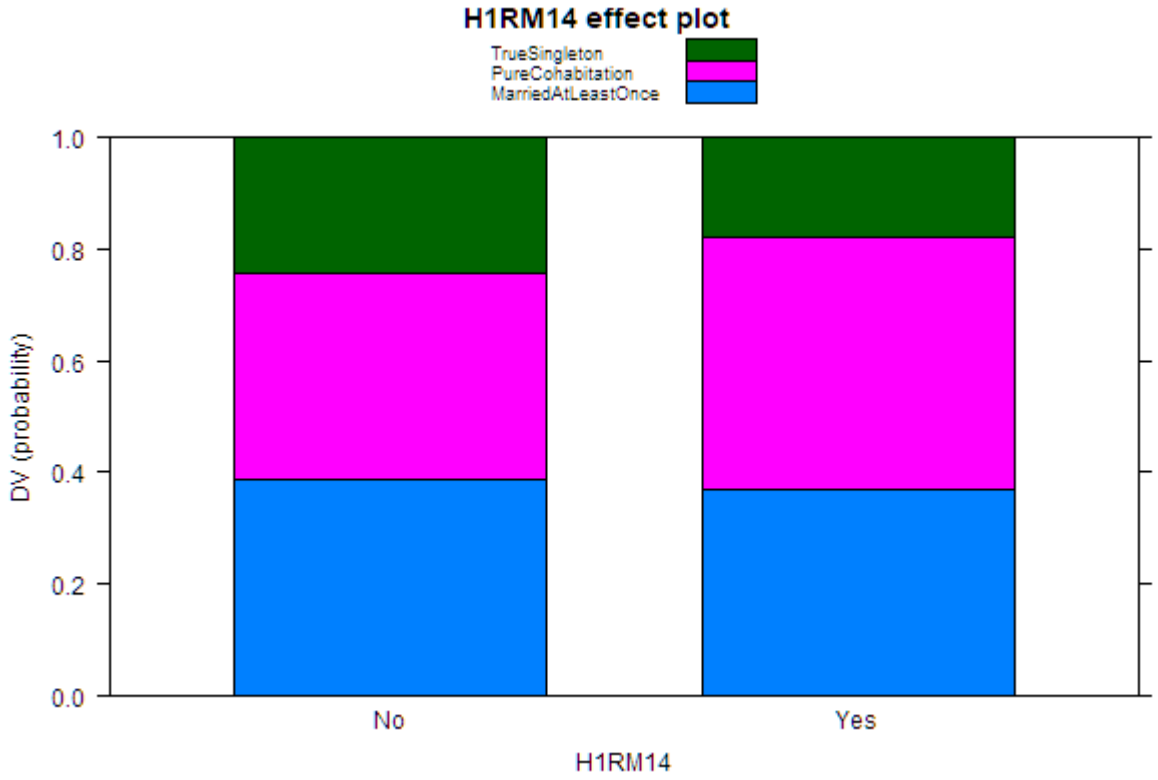
Note. H3OD4B (Wave III): "What is your race (check all that apply): black or African American"



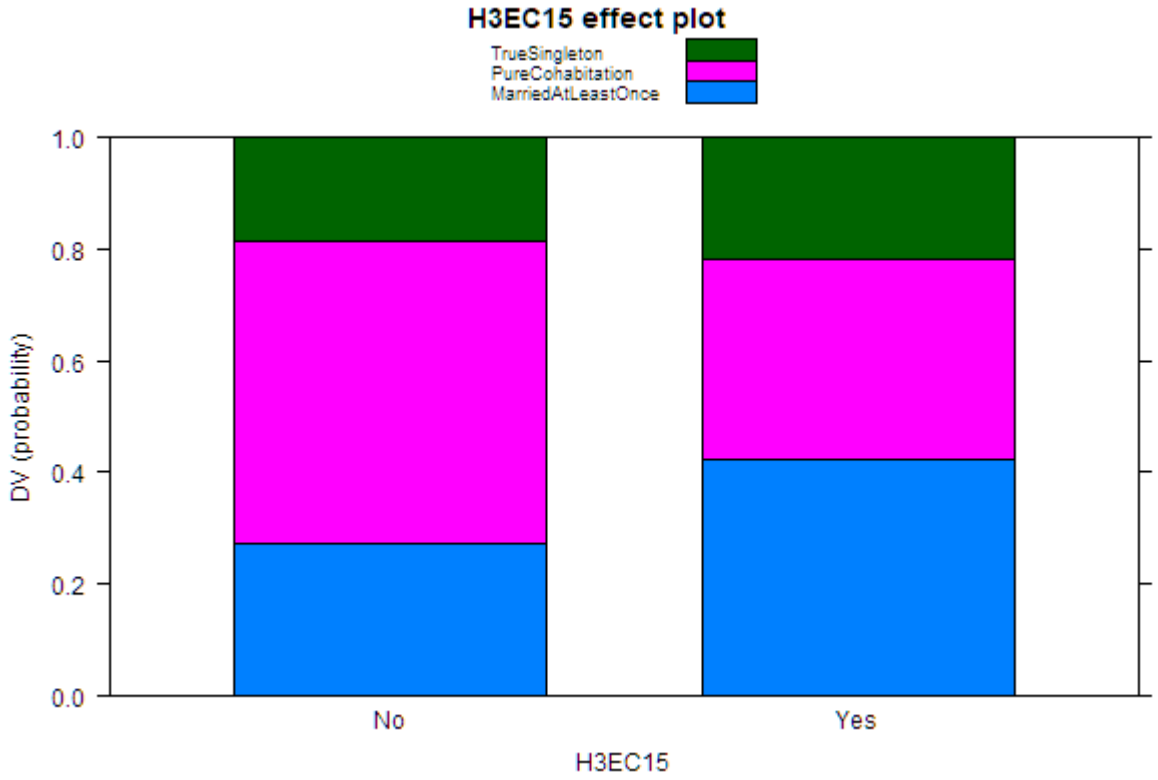
Note. H3SE4 (Wave III): "With how many different partners have you had vaginal intercourse in the past 12 months?"



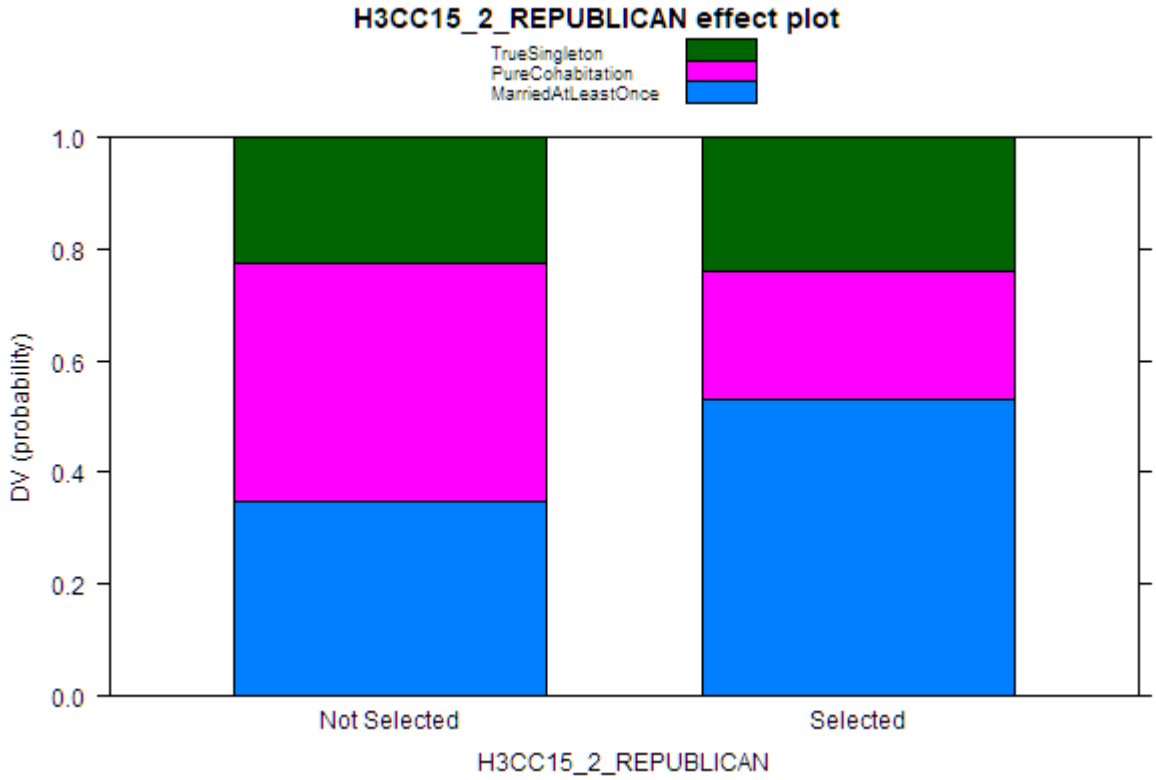
Note. H3SE7 (Wave III): "On how many of these occasions of vaginal intercourse in the past 12 months did you or your partner use some form of birth control or pregnancy protection?" Response: 4 - all



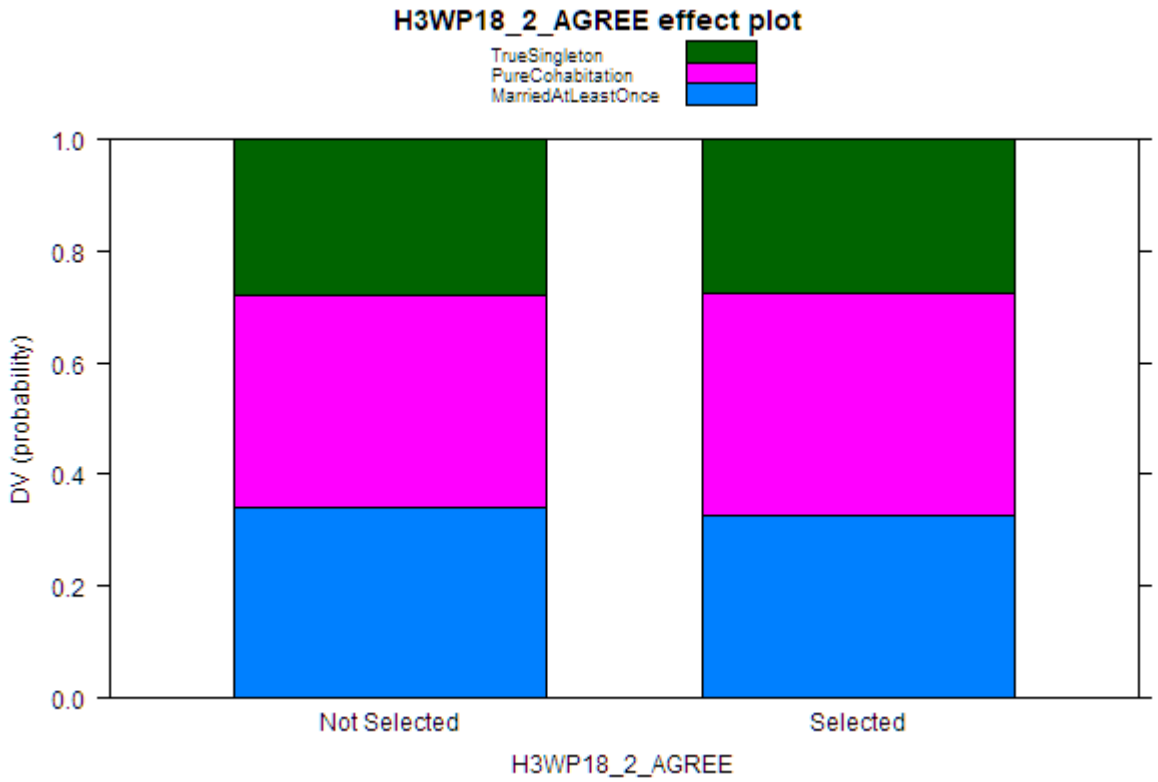
Note. H1RM14 (Wave I): "Has she [resident mother] ever smoked cigarettes?"



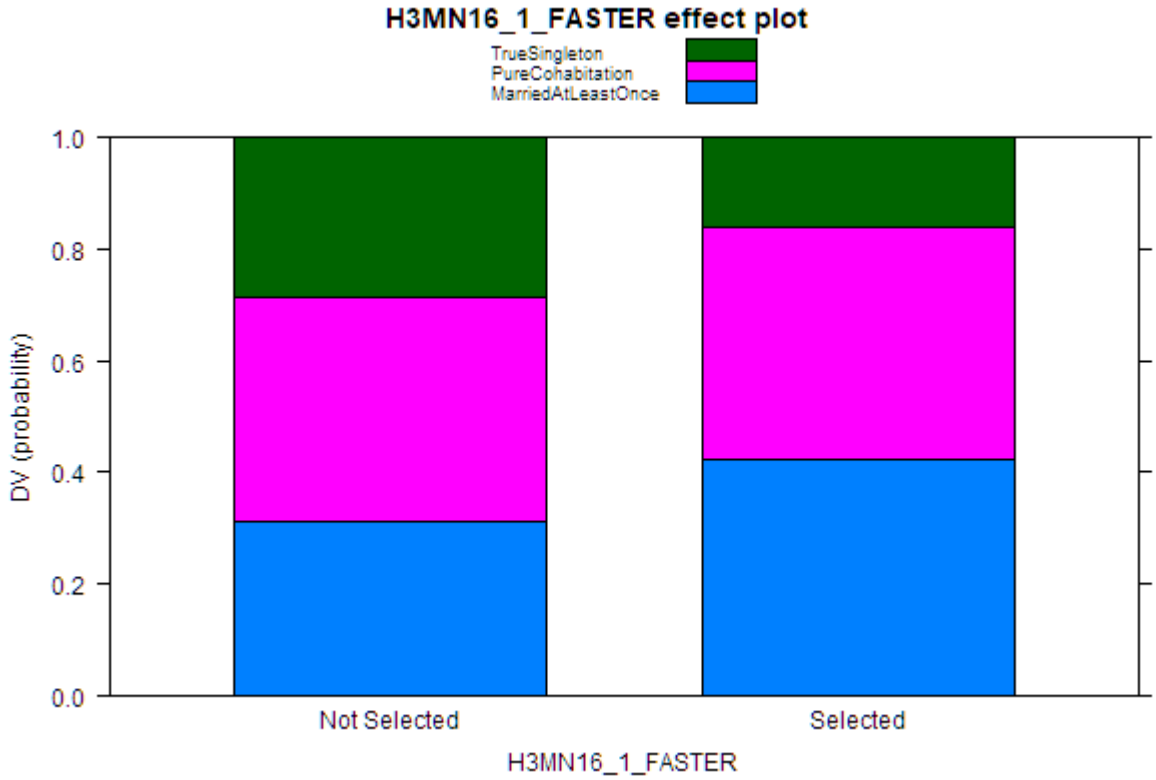
Note. H3EC15 (Wave III): "Do you have a checking account?"



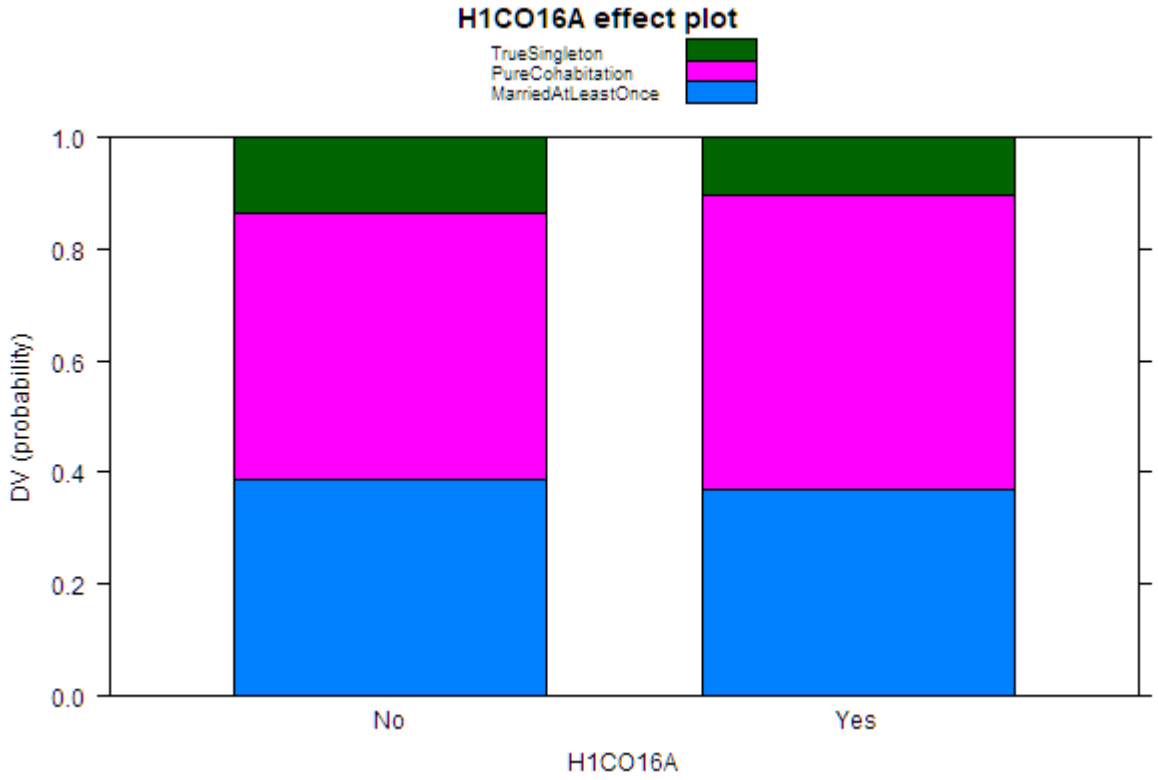
Note. H3CC15 (Wave III): "With which [political] party do you identify?" Response: 2 - Republican



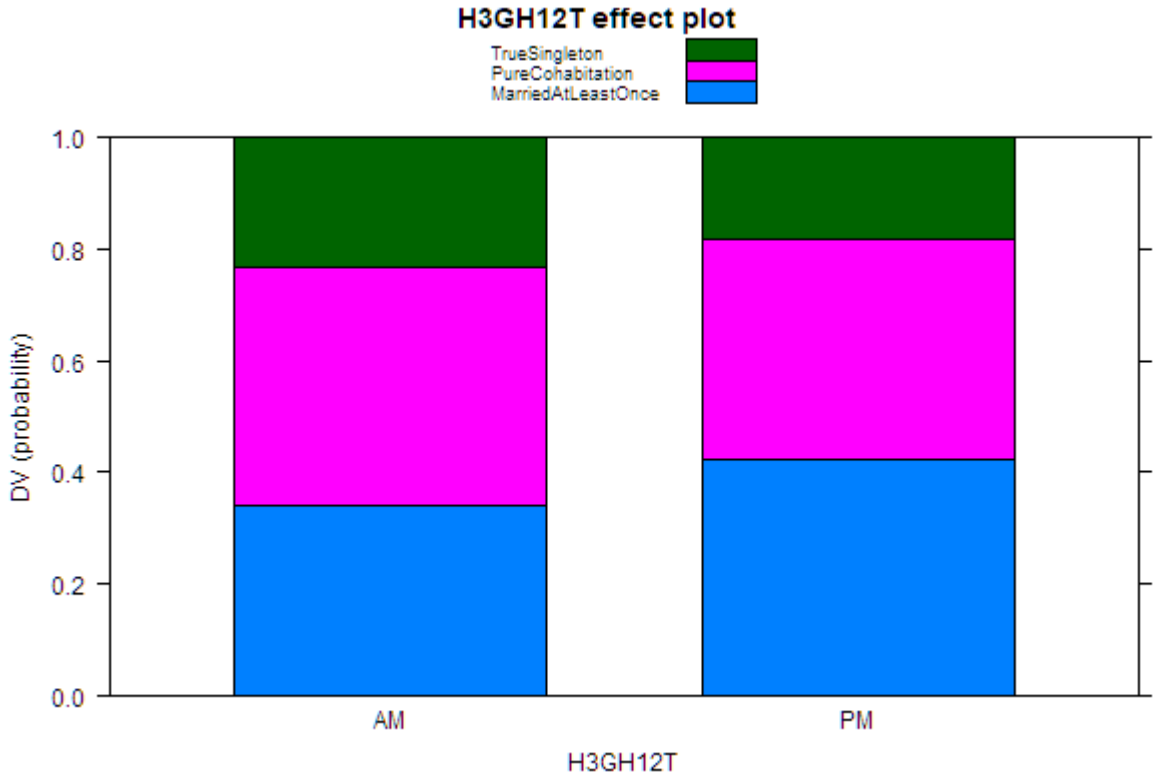
Note. H3WP18 (Wave III): "How much do you agree or disagree with the next statement? You enjoy doing things with your [current residential] mother." Response: 2 - agree



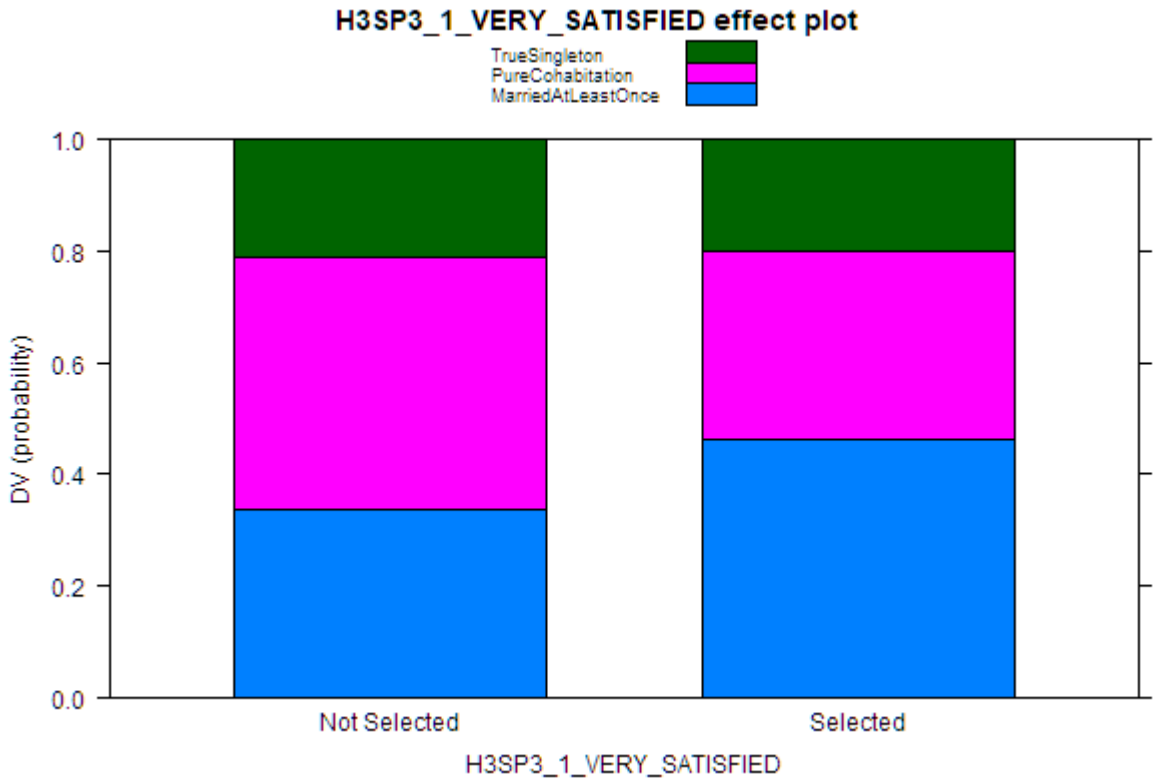
Note. H3MN16 (Wave III): "In terms of taking on adult responsibilities, would you say you grew up faster, slower, or at about the same rate?" Response: 1 - faster



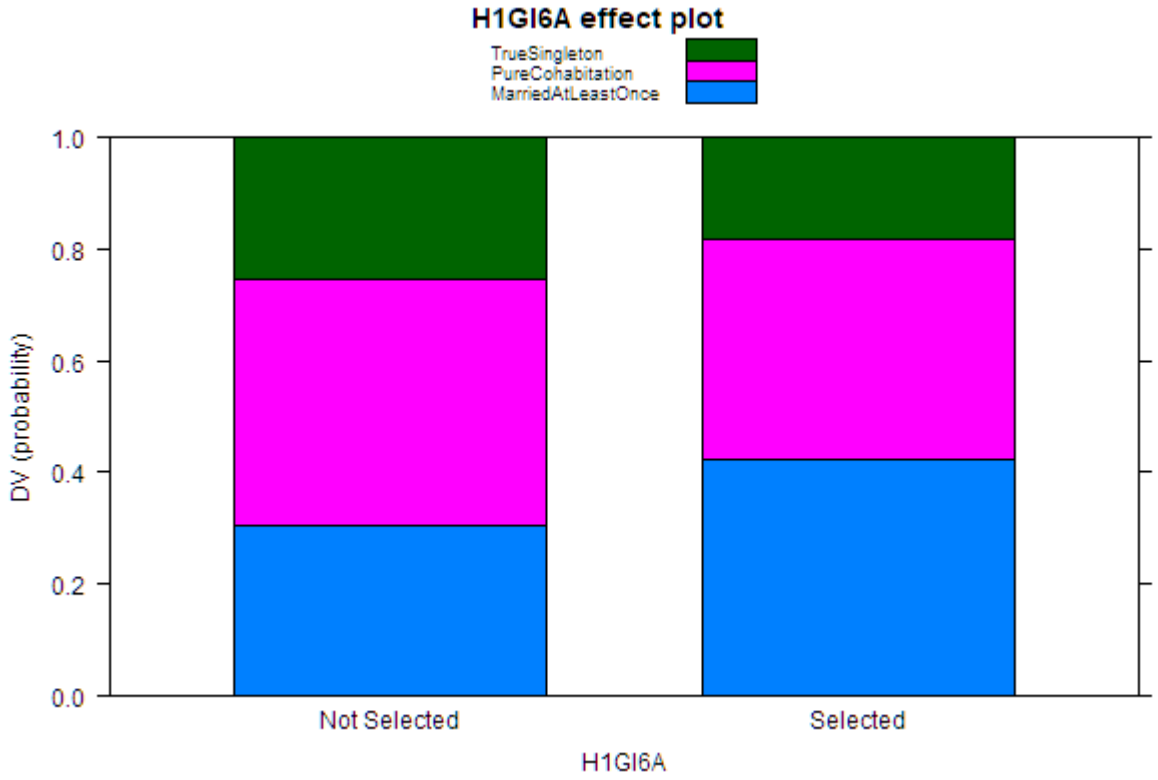
Note. H1CO16A (Wave I): "Have you ever been told by a doctor or a nurse that you had Chlamydia?"



Note. H3GH12T (Wave III): "On days when you go to work, school, or similar activities, what time do you usually go to sleep the night (or day) before?"

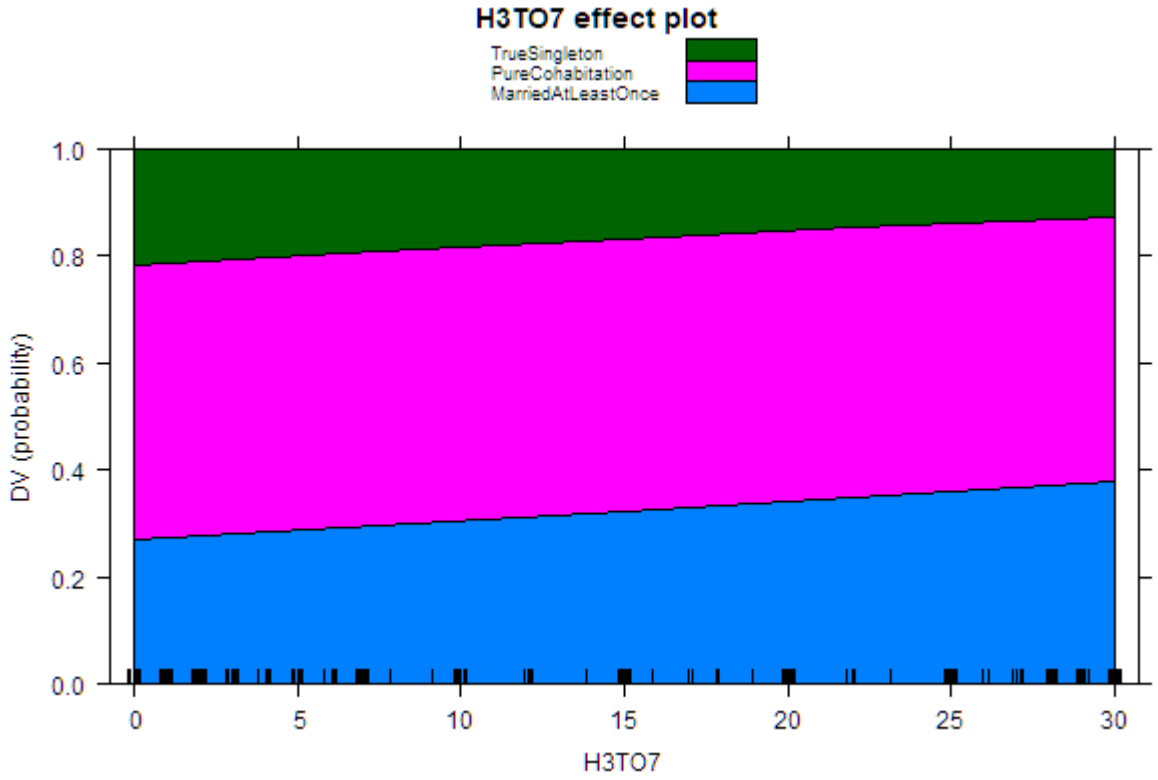


Note. H3SP3 (Wave III): "How satisfied are you with your life as a whole?" Response: 1 - very satisfied

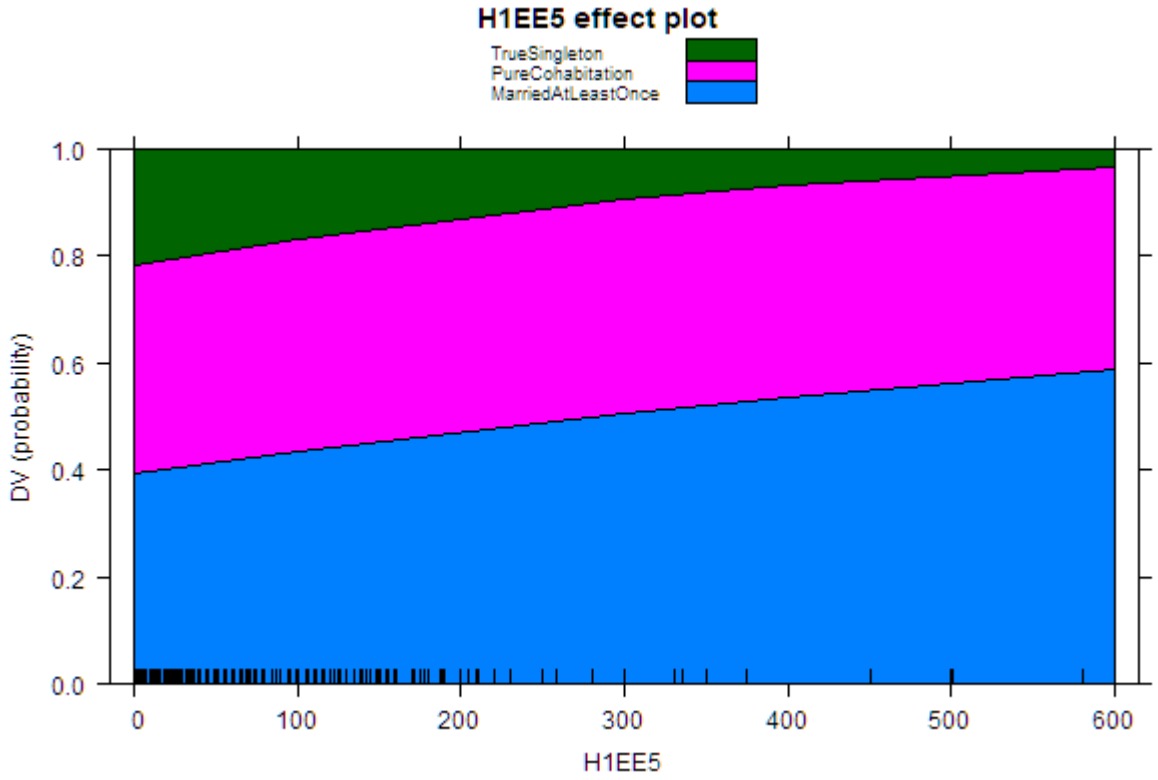


Note. H1GI6A (Wave I): "What is your race (check all that apply): white"

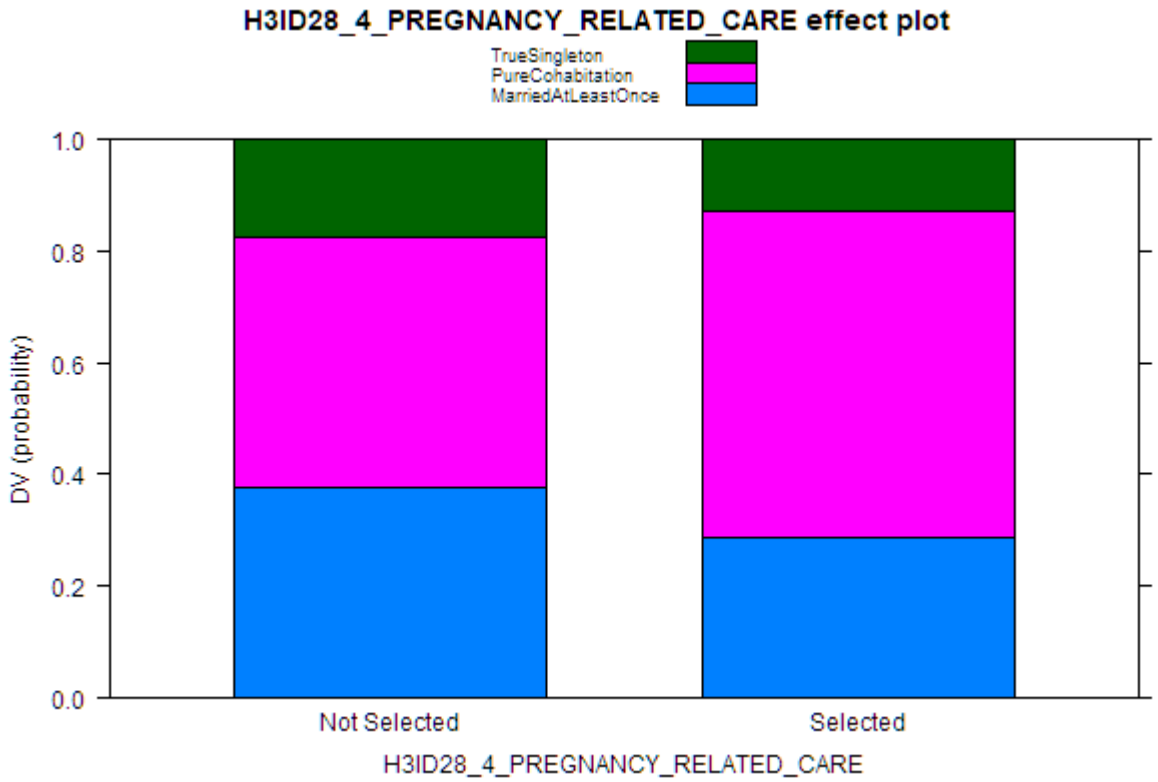
Note. H3RE41 (Wave III): "To what extent are you a religious person?" Response: 3 - very religious



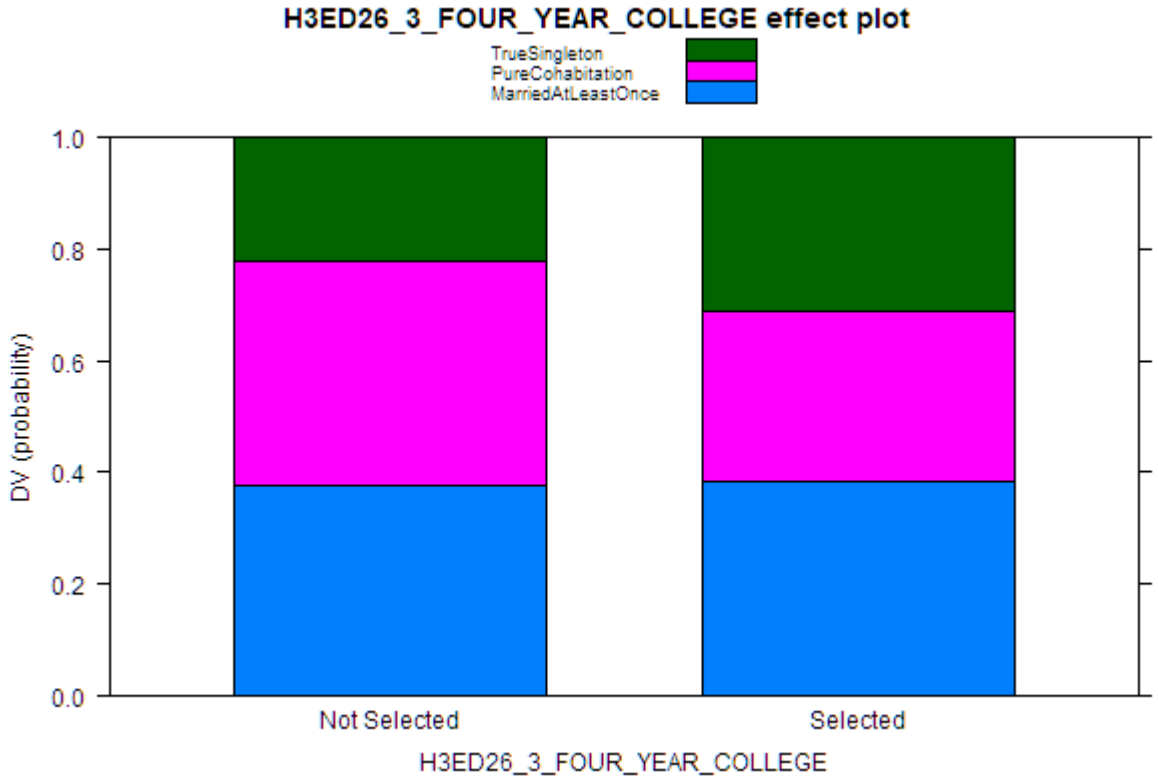
Note. H3TO7 (Wave III): "During the past 30 days, on how many days did you smoke cigarettes?"



Note. H1EE5 (Wave I): "How much money do you earn in a typical non-summer week from all your jobs combined?"

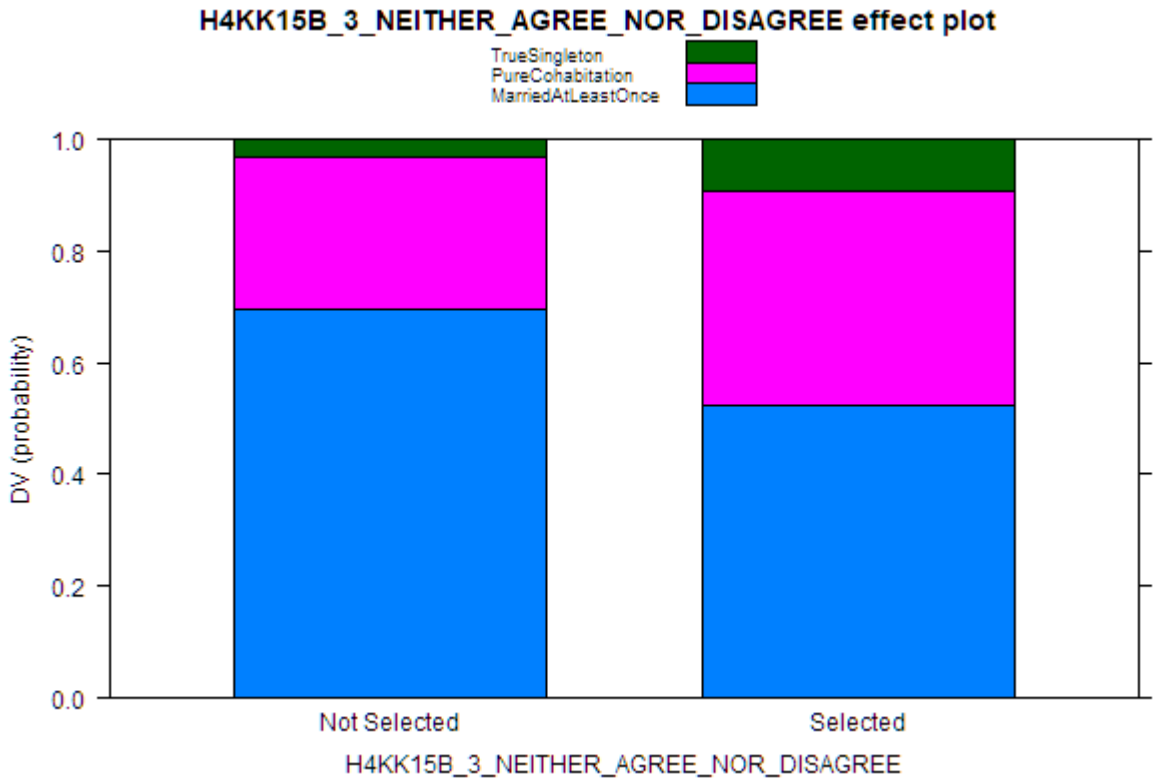


Note. H3ID28 (Wave III): "What was the main reason for your most recent emergency room visit?" Response: 4 - pregnancy-related care

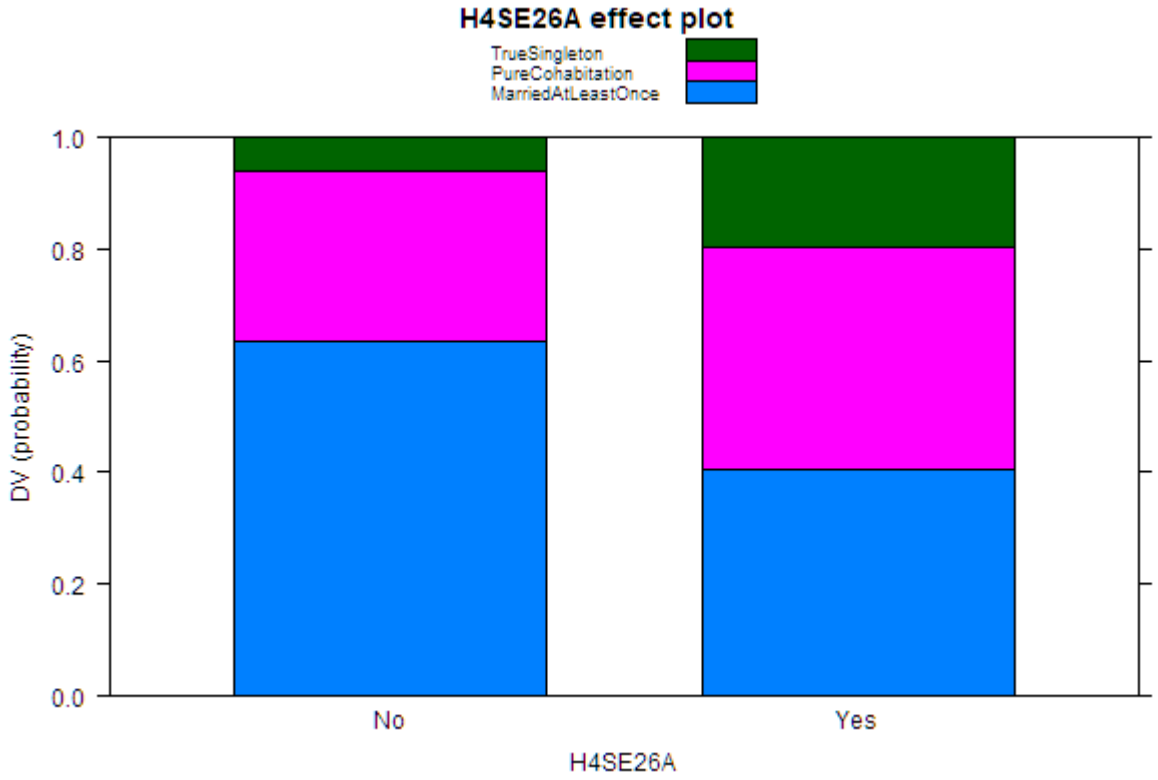


Note. H3ED26 (Wave III): "Is this a high school, a two-year college, a four-year college, or a graduate school?" Response: 3 - four year college

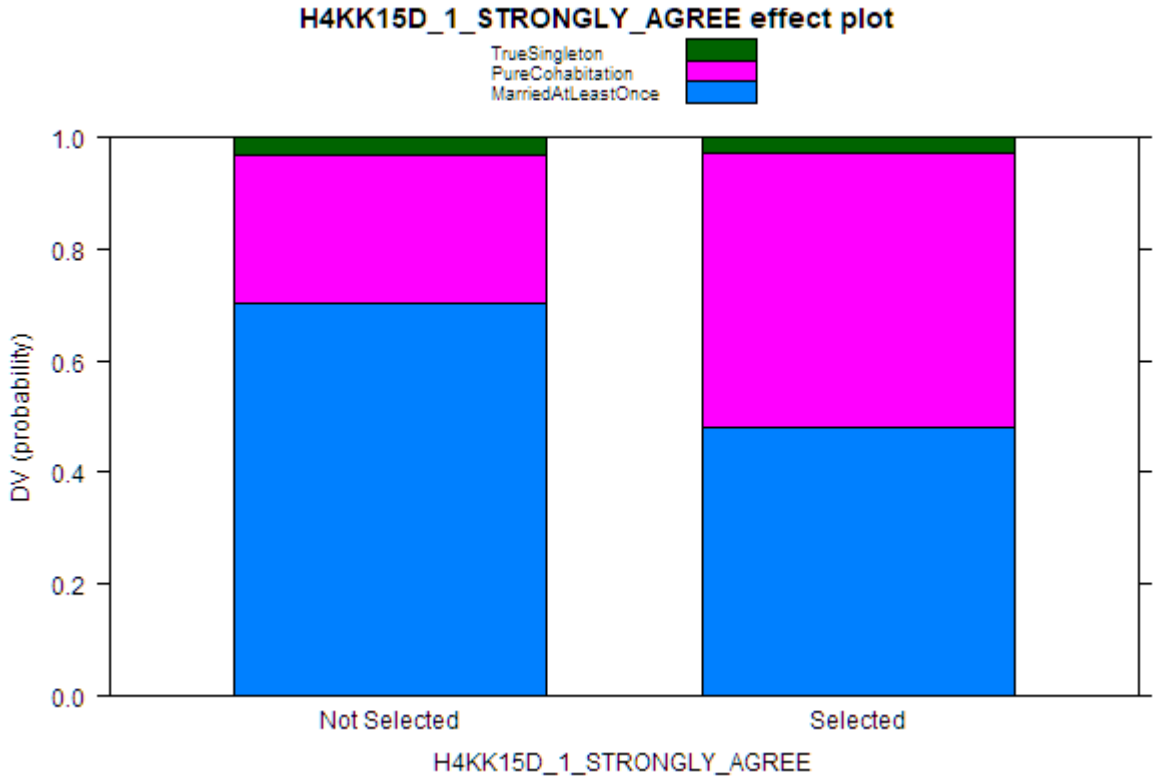
Appendix D



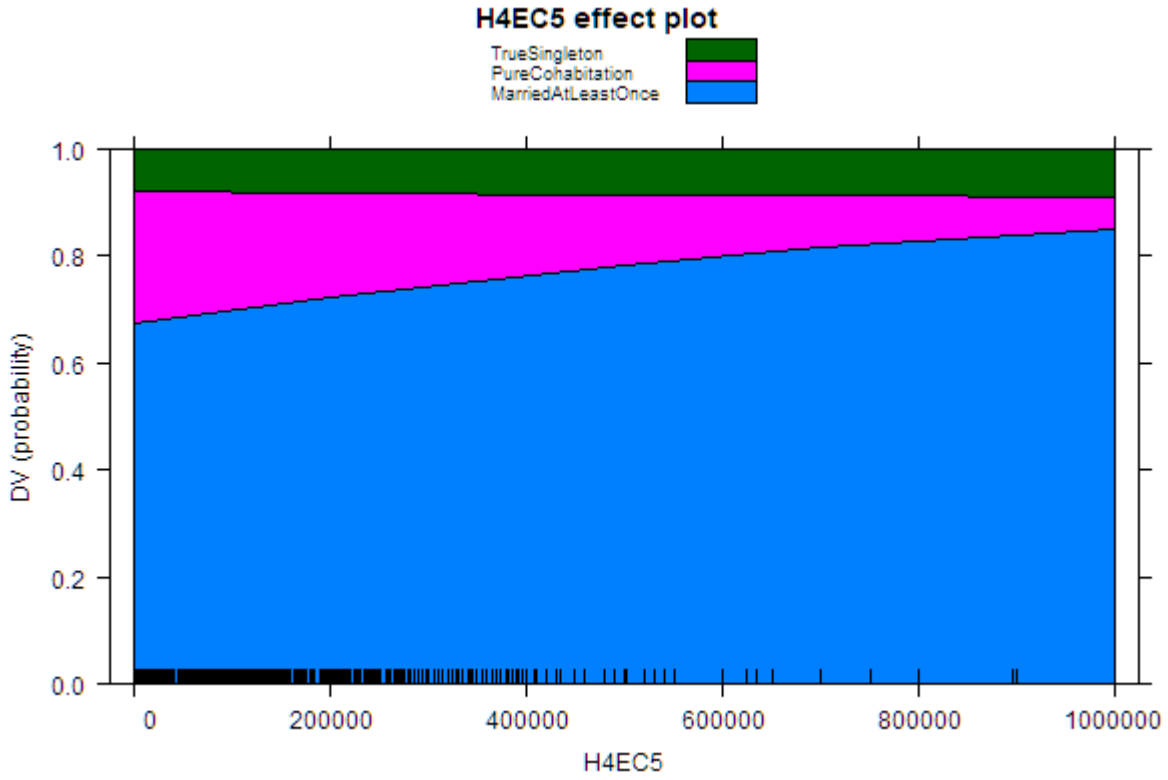
Note. H4KK15B (Wave IV): "How much do you agree or disagree with the following statement? I feel close to my child(ren)." Response: 3 - neither agree nor disagree



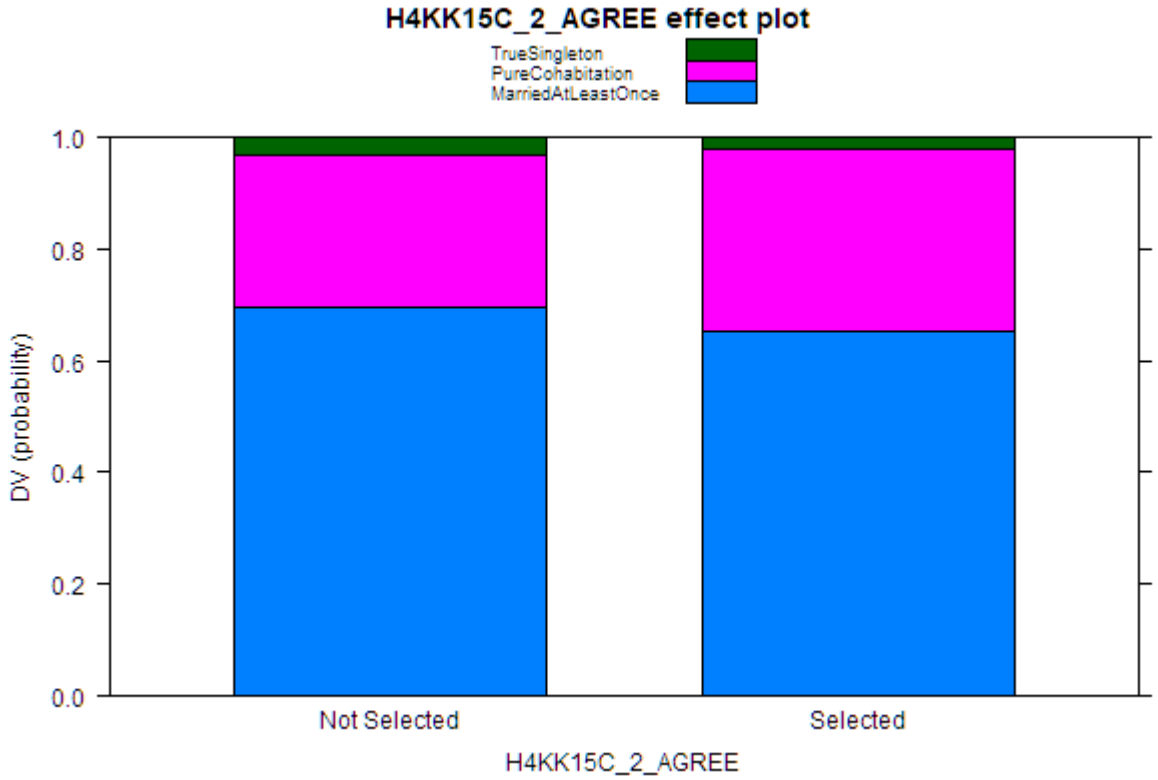
Note. H4SE26A (Wave IV): "In the past 12 months, did you or your partner(s) use any of these methods for birth control or disease prevention (check all that apply): condoms (rubbers)"



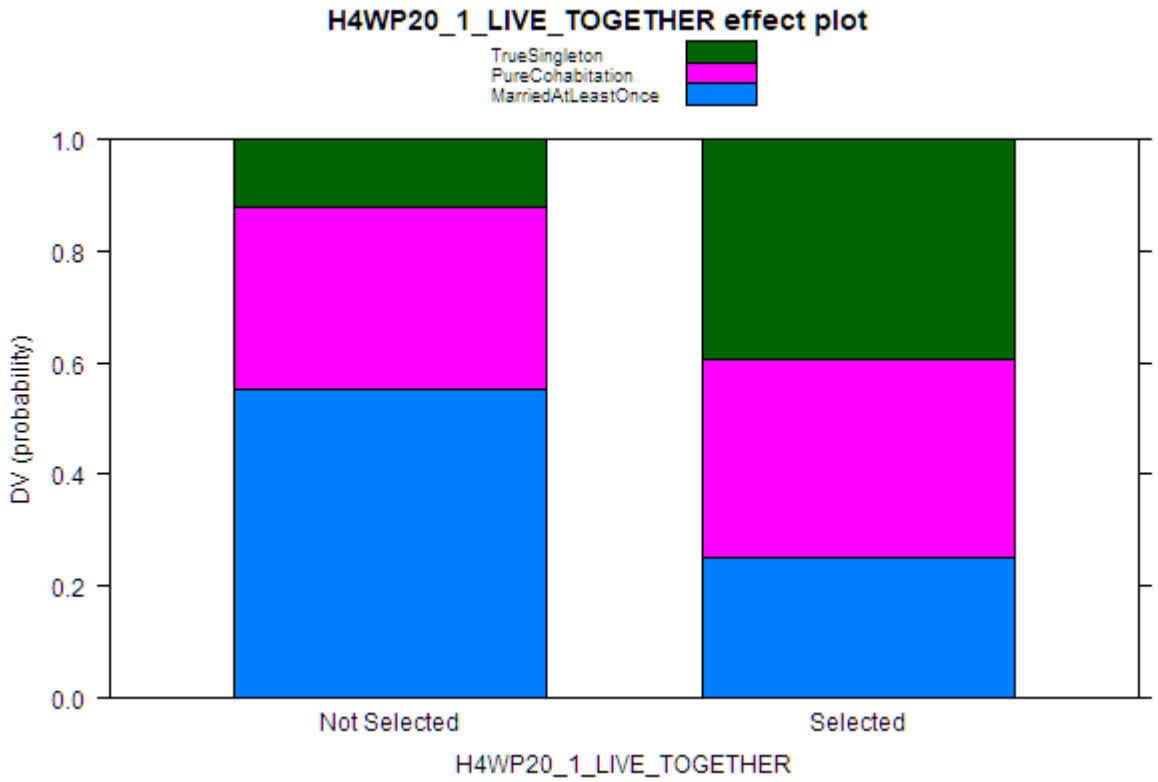
Note. H4KK15D (Wave IV): "How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being a parent." Response: 1 - strongly agree



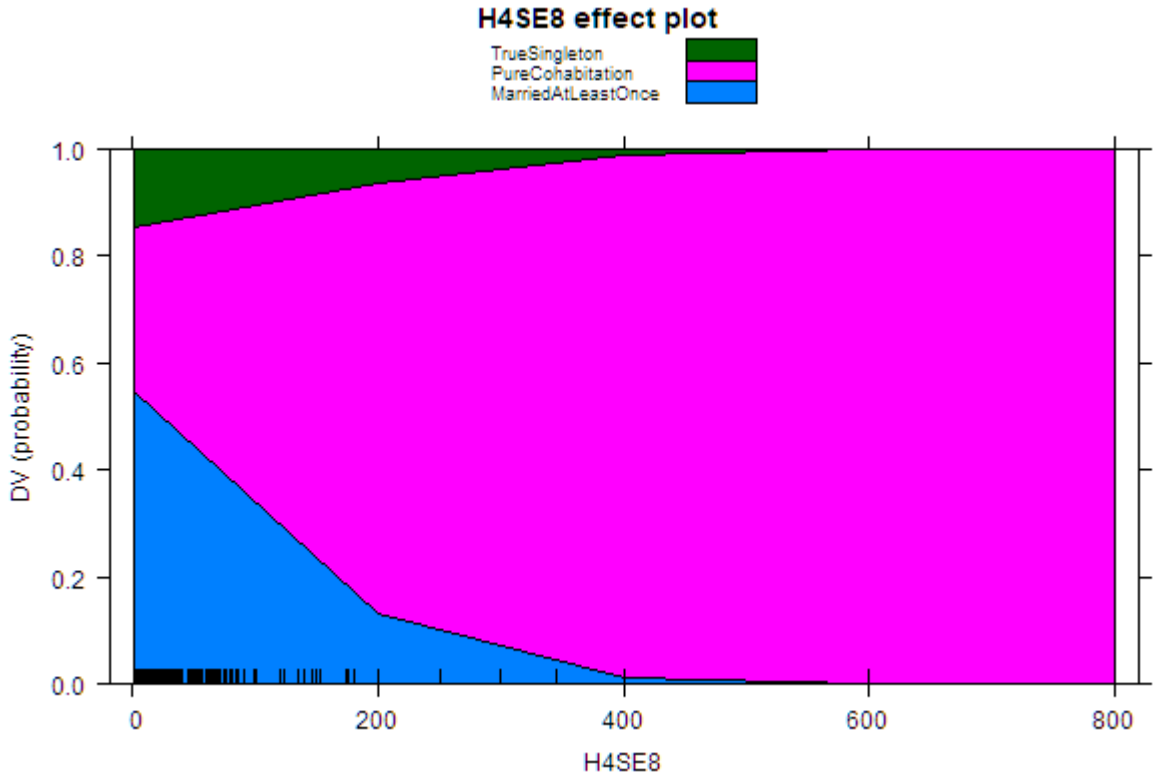
Note. H4EC5 (Wave IV): "About how much do {YOU AND/OR YOUR SPOUSE/PARTNER} owe on the mortgage for your house, apartment, or residence?"



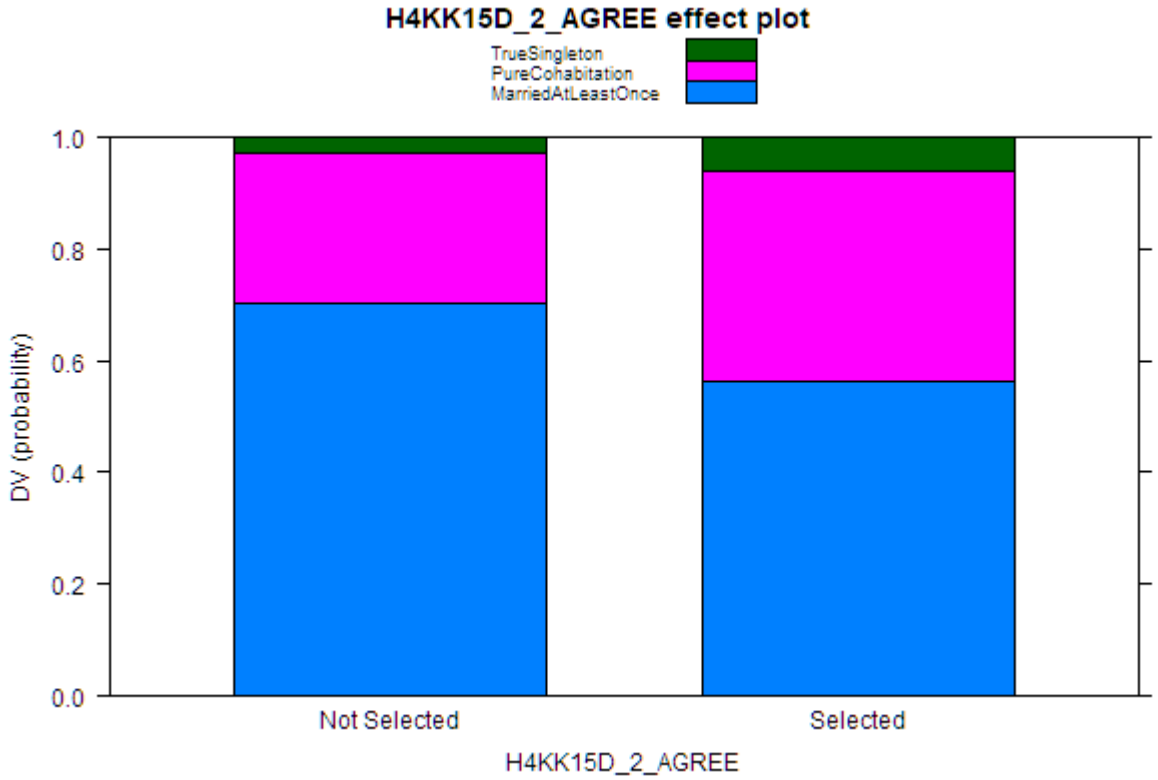
Note. H4KK15C (Wave IV): "How much do you agree or disagree with the following statement? The major source of stress in my life is my child(ren)." Response: 2 - agree



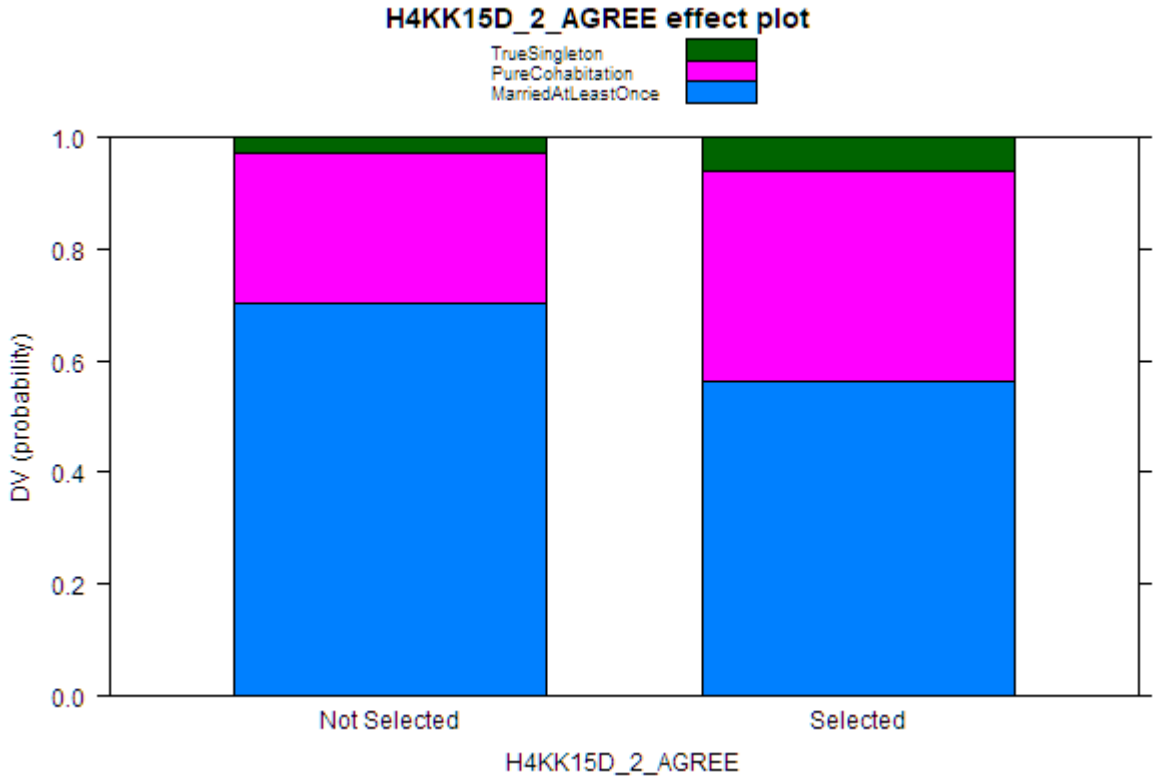
Note. H4WP20 (Wave IV): "How far do you and your [mother figure] live from one another?" Response: 1 - live together



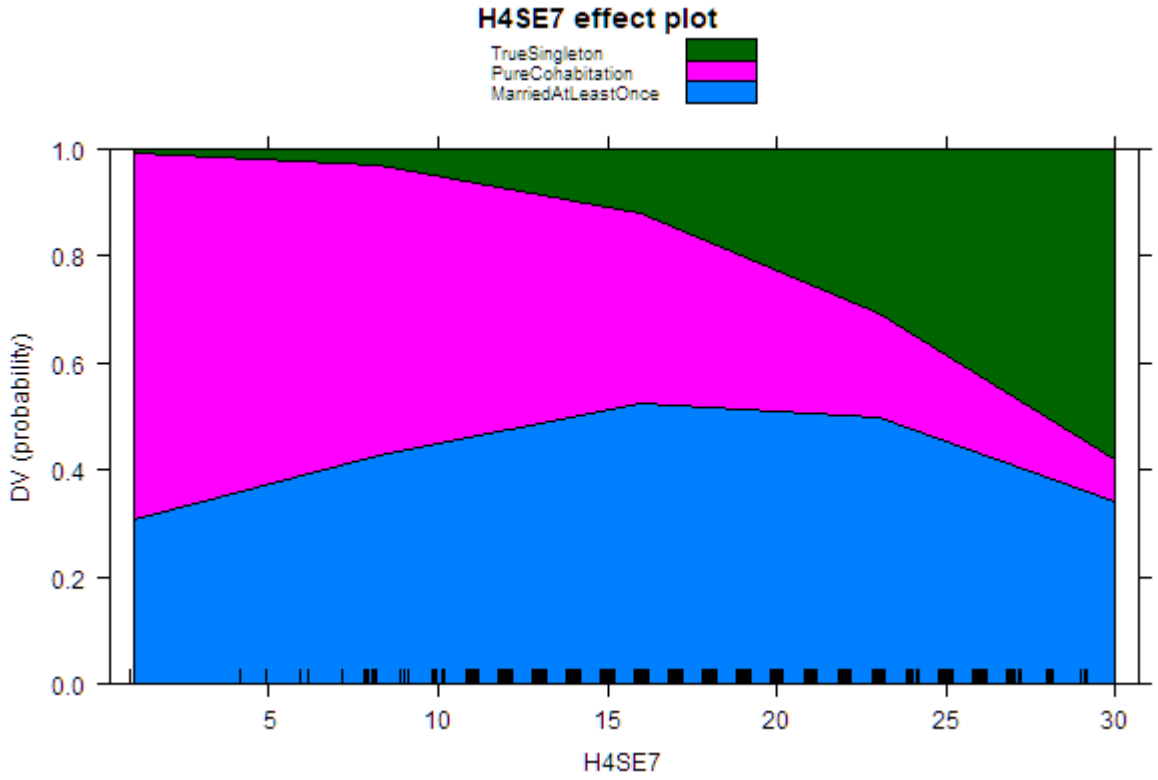
Note. H4SE8 (Wave IV): "With how many partners have you ever had vaginal intercourse, even if only once?"



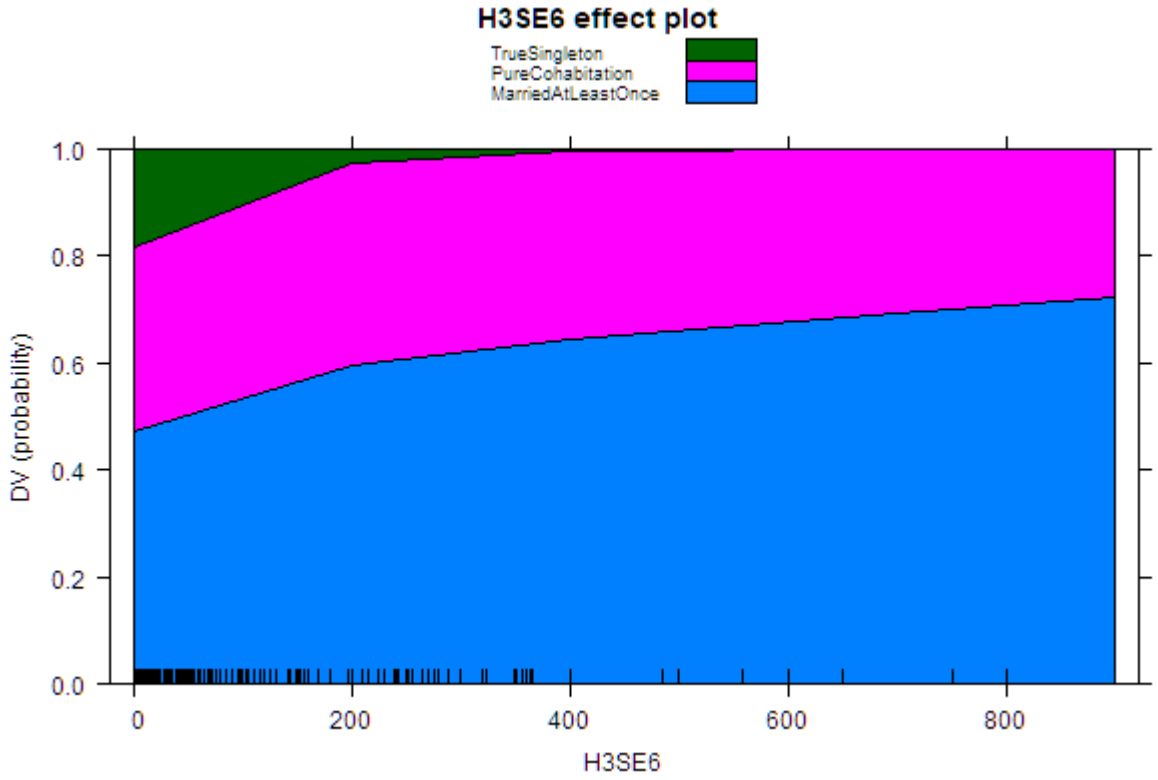
Note. H4KK15D (Wave IV): "How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being a parent." Response: 2 - agree



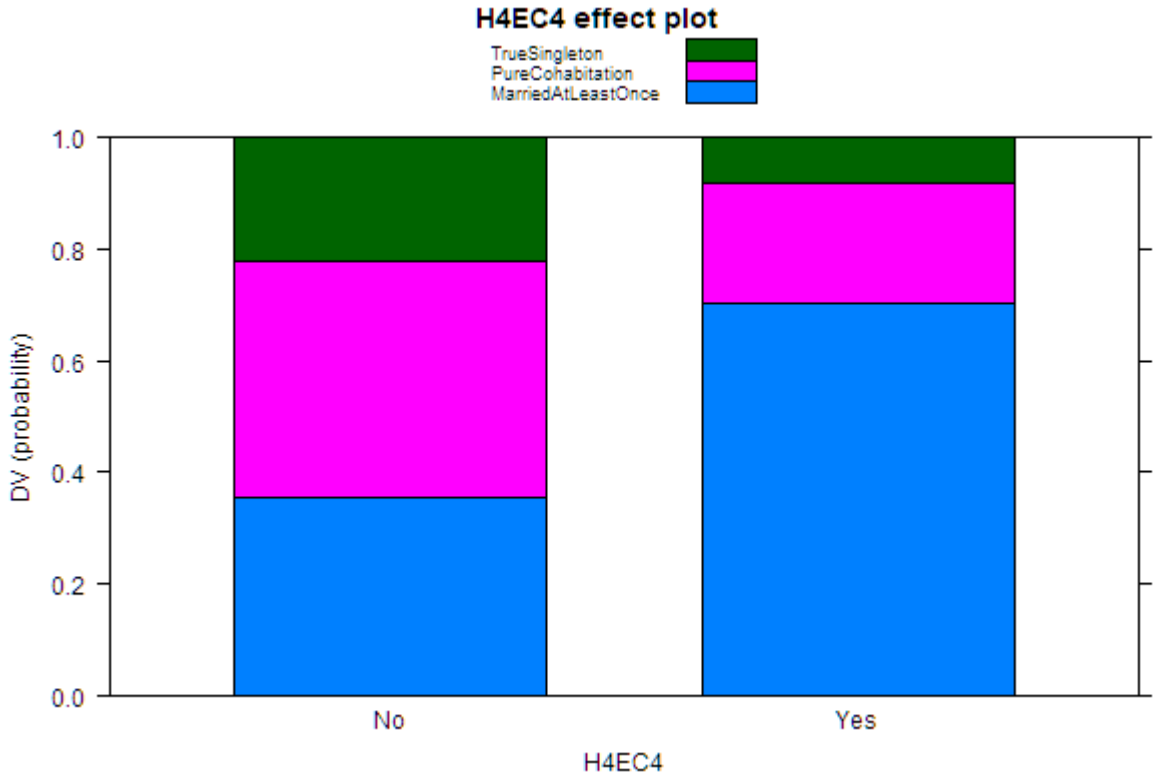
Note. H4KK15D (Wave IV): "How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being a parent." Response: 2 - agree



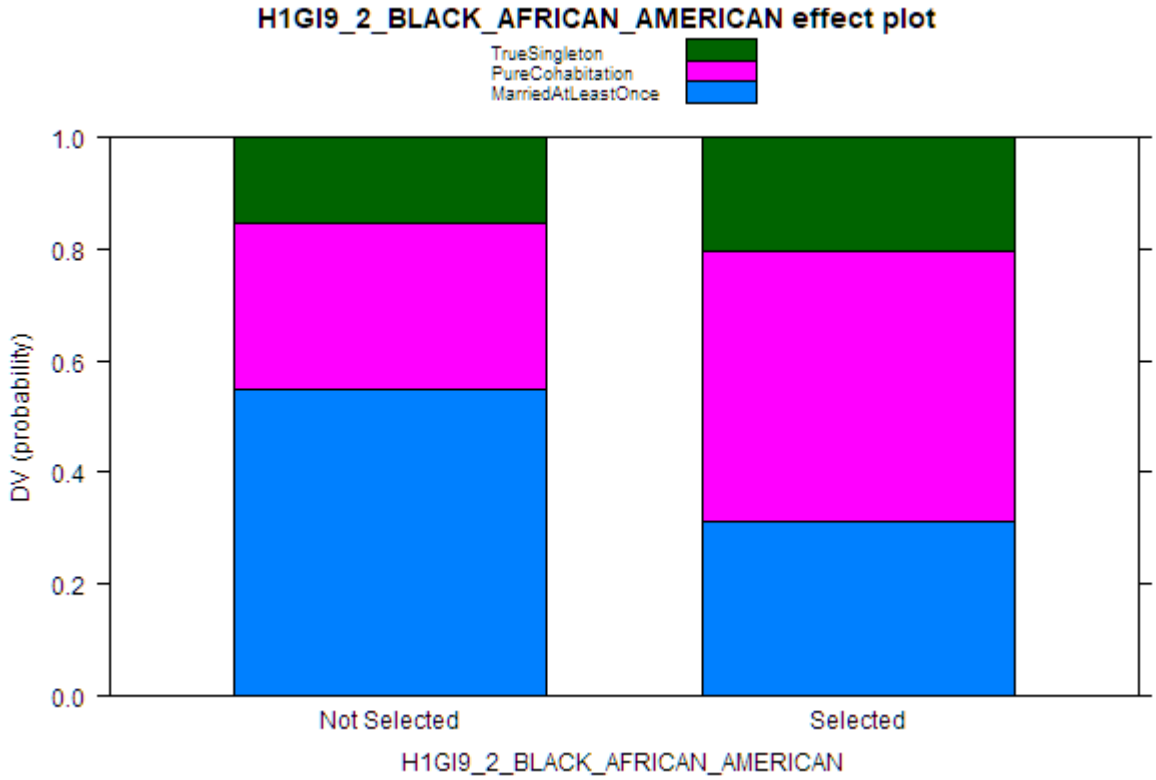
Note. H4SE7 (Wave IV): "How old were you the first time you ever had vaginal intercourse?"



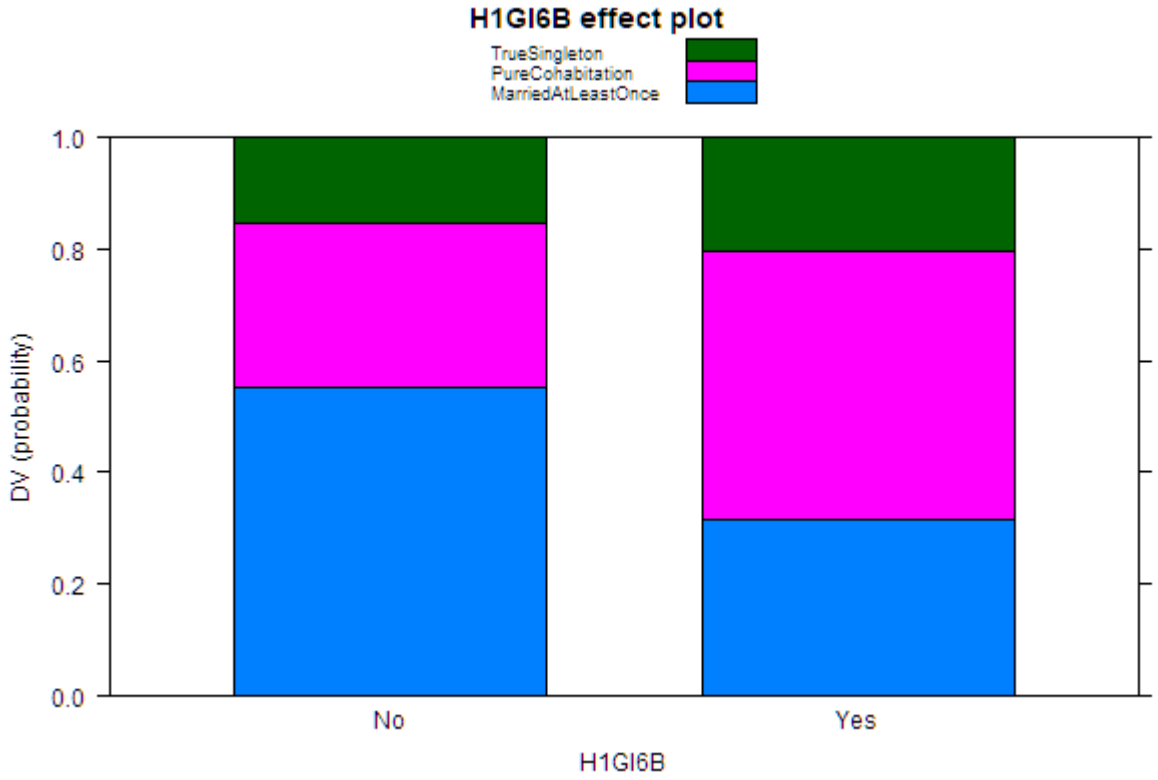
Note. H3SE6 (Wave III): "How many times have you had vaginal intercourse in the past 12 months?"



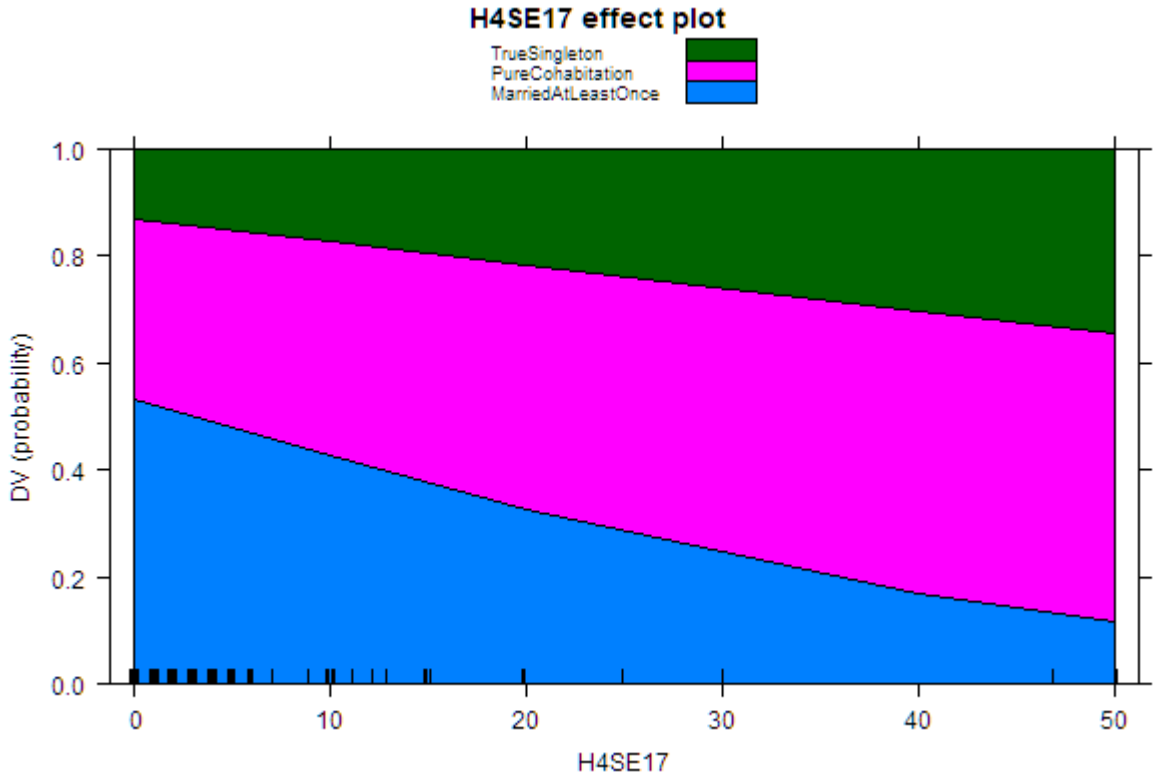
Note. H4EC4 (Wave IV): "Is your house, apartment, or residence owned or being bought by {YOU AND/OR YOUR SPOUSE/PARTNER}?"



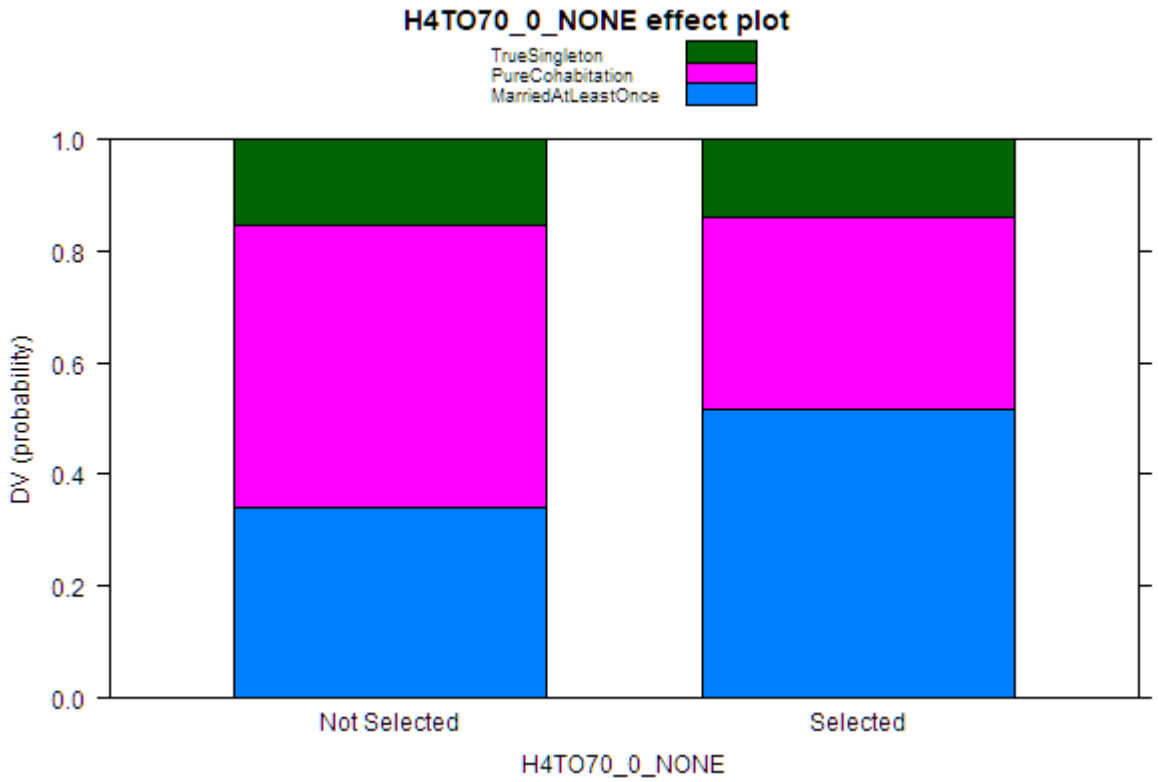
Note. H1GI9 (Wave I): "Interviewer: Please code the race of the respondent from your observation alone." Response: 2 - Black or African American



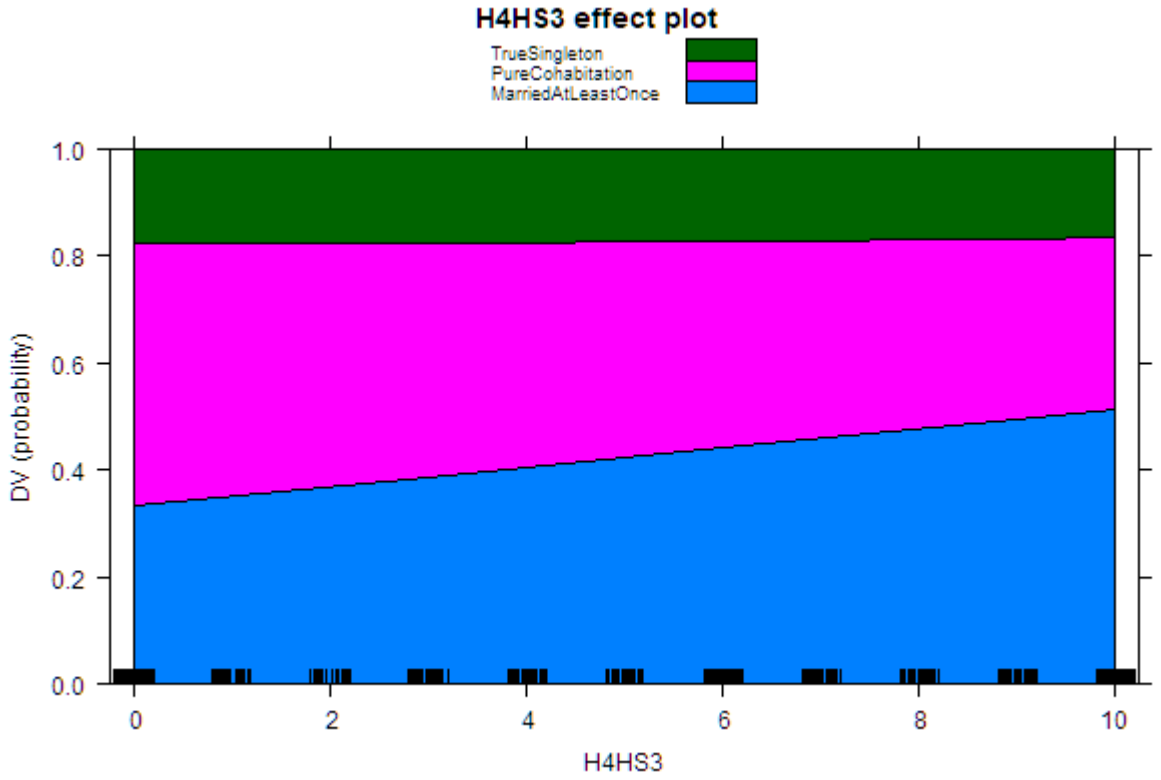
Note. H1GI6B (Wave I): "What is your race (check all that apply): black or African American"



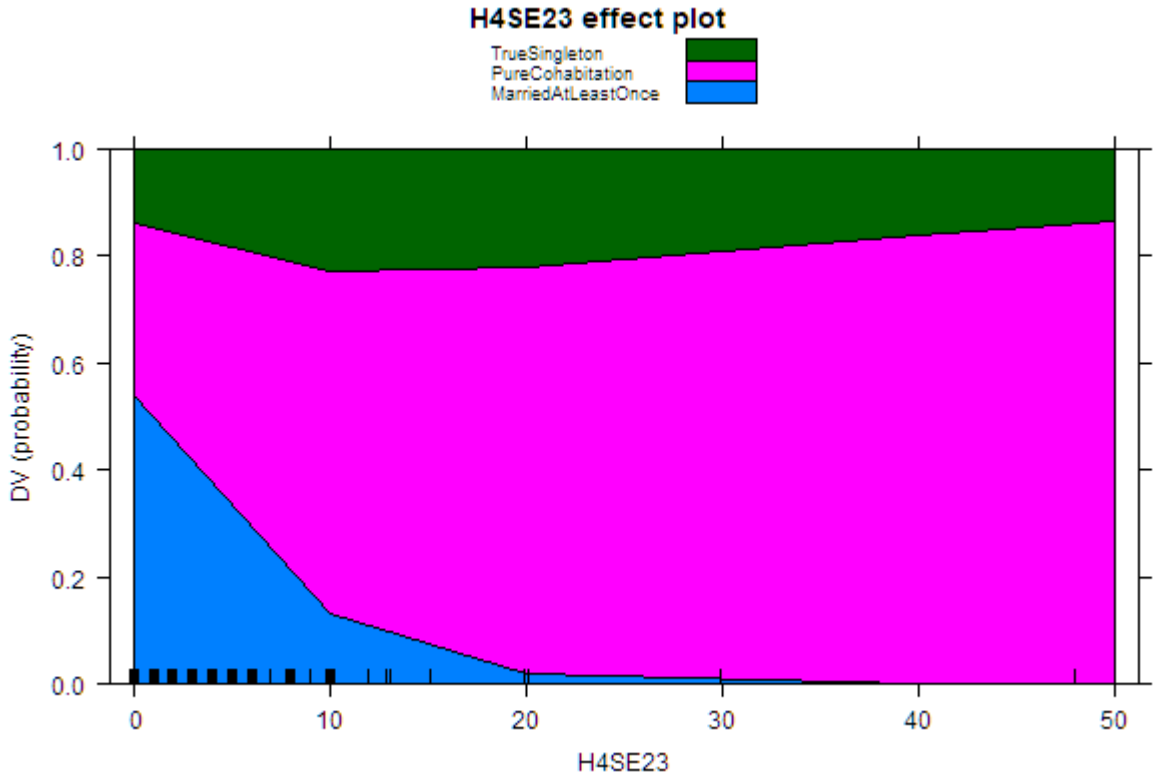
Note. H4SE17 (Wave IV): "Considering all types of sexual activity, with how many male partners have you had sex in the past 12 months, even if only one time?"



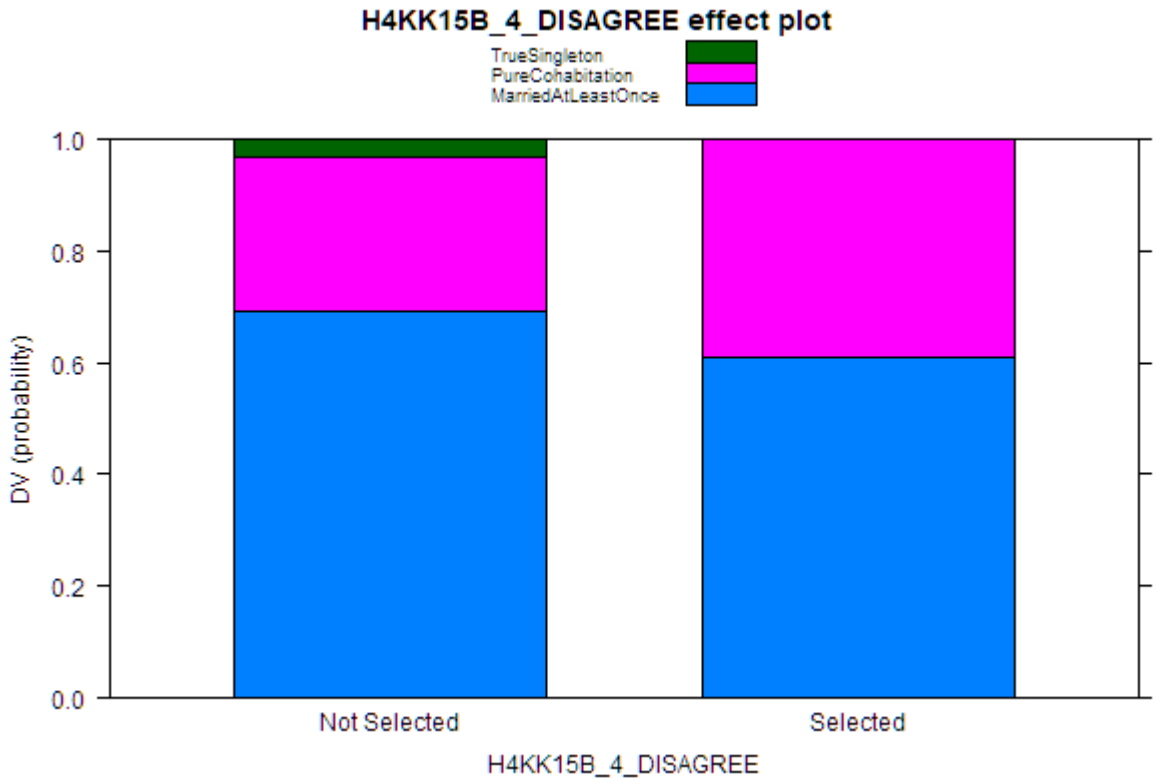
Note. H4TO70 (Wave IV): "During the past 12 months, on how many days did you use marijuana?" Response: 0 - none



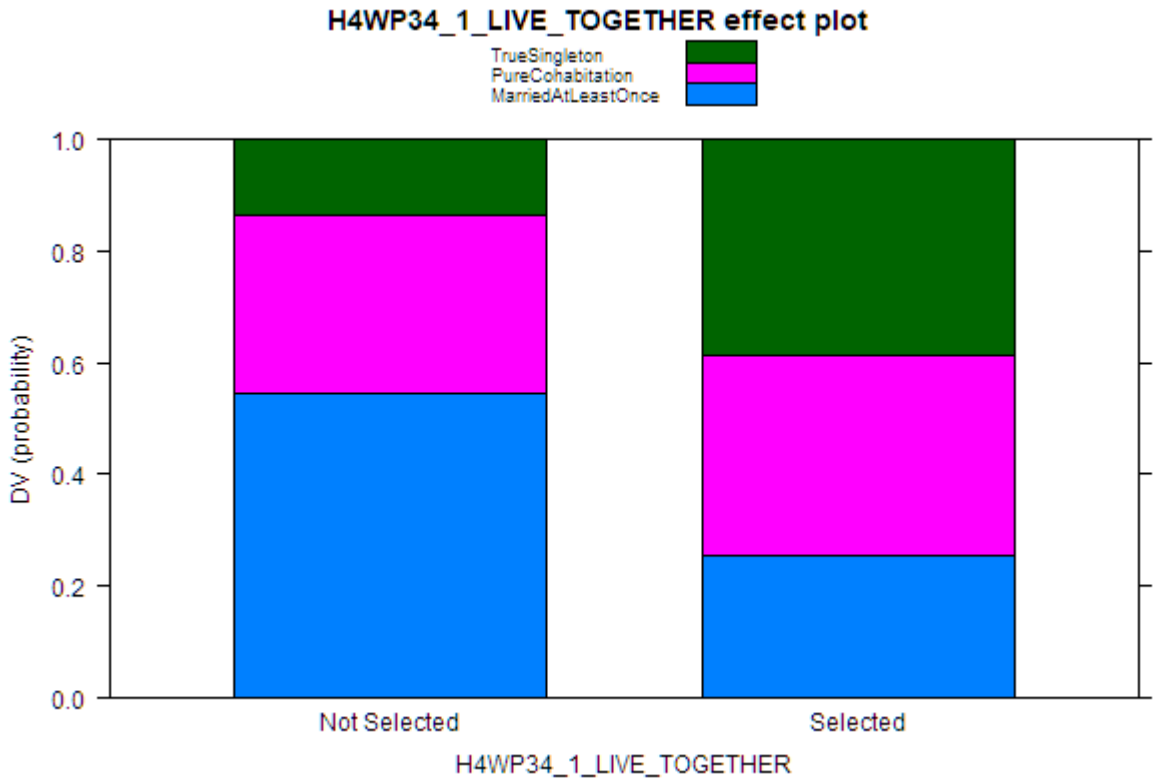
Note. H4HS3 (Wave IV): "Over the past 12 months, how many months did you have health insurance?"



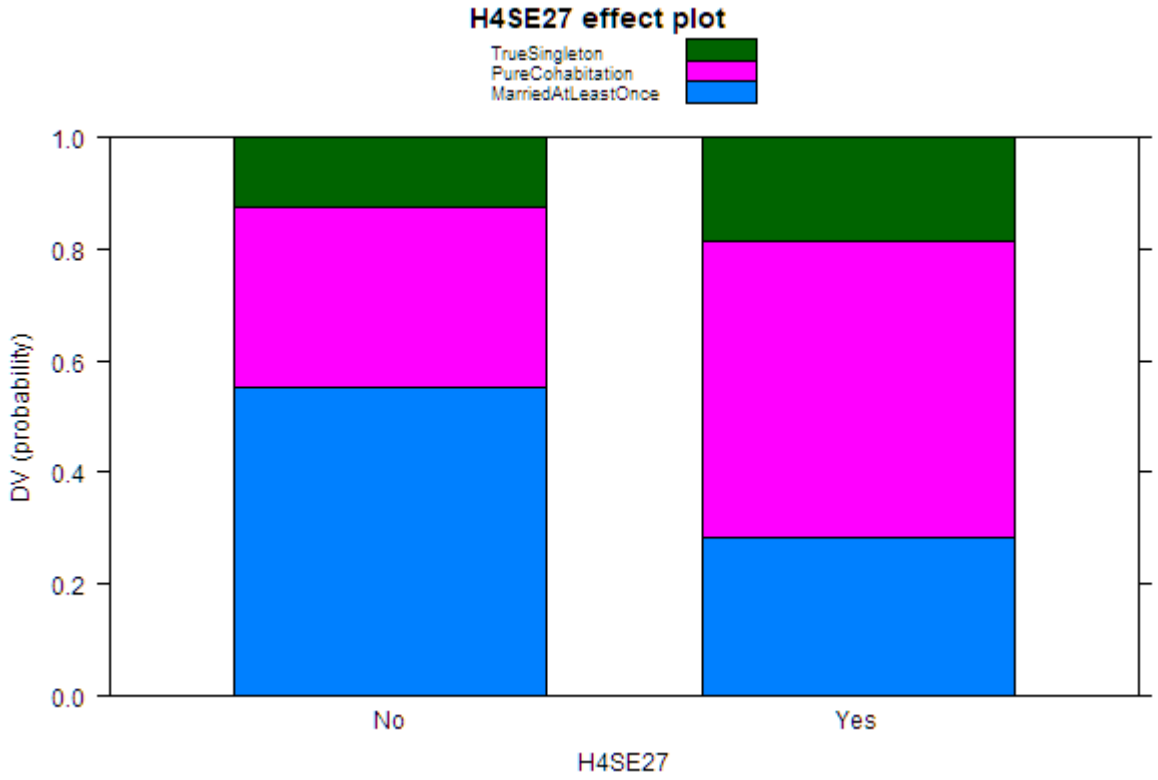
Note. H4SE23 (Wave IV): "Considering all types of sexual activity, with how many female partners have you had sex in the past 12 months?"



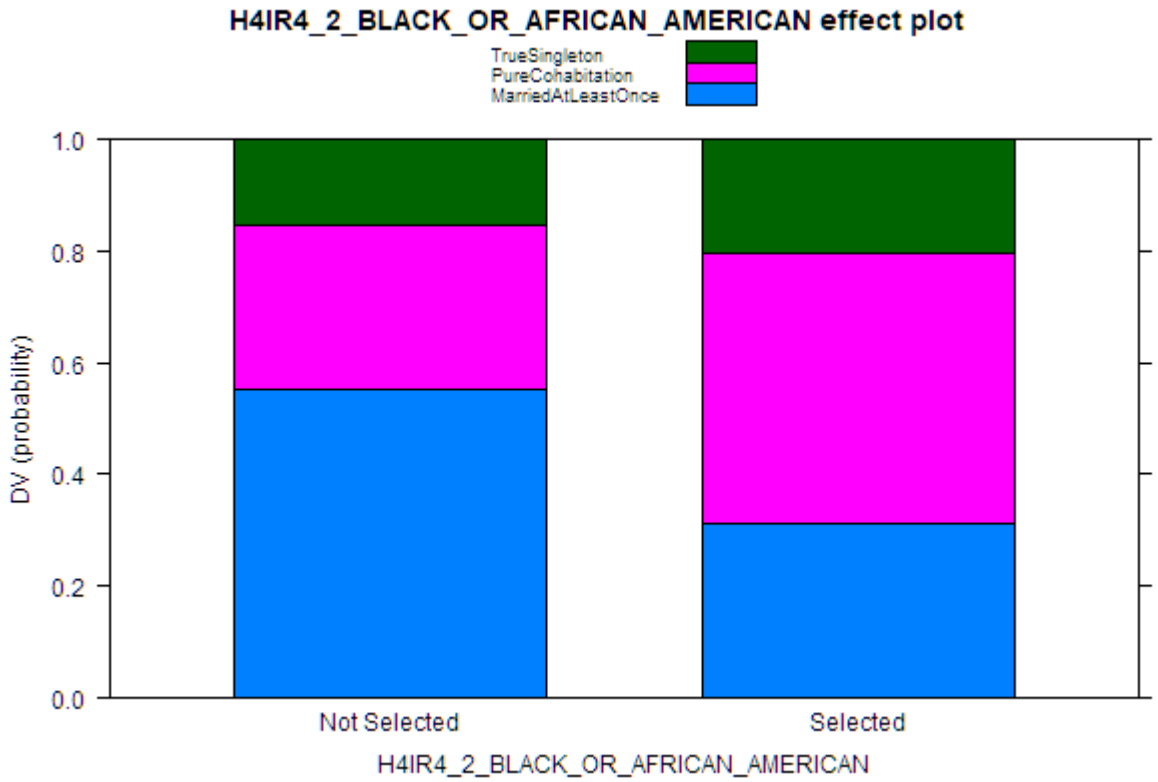
Note. H4KK15B (Wave IV): "How much do you agree or disagree with the following statement? I feel close to my child(ren)." Response: 4 - disagree



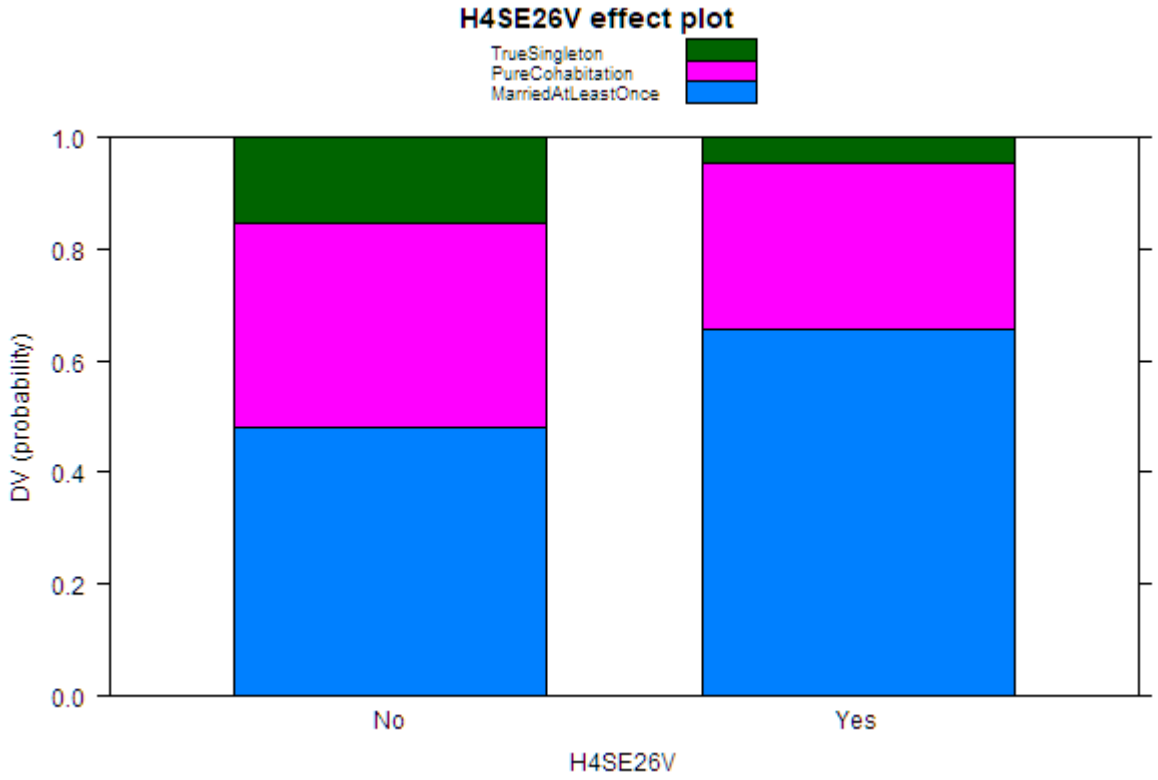
Note. H4WP34 (Wave IV): "How far do you and your [father figure] live from one another?" Response: 1 - live together



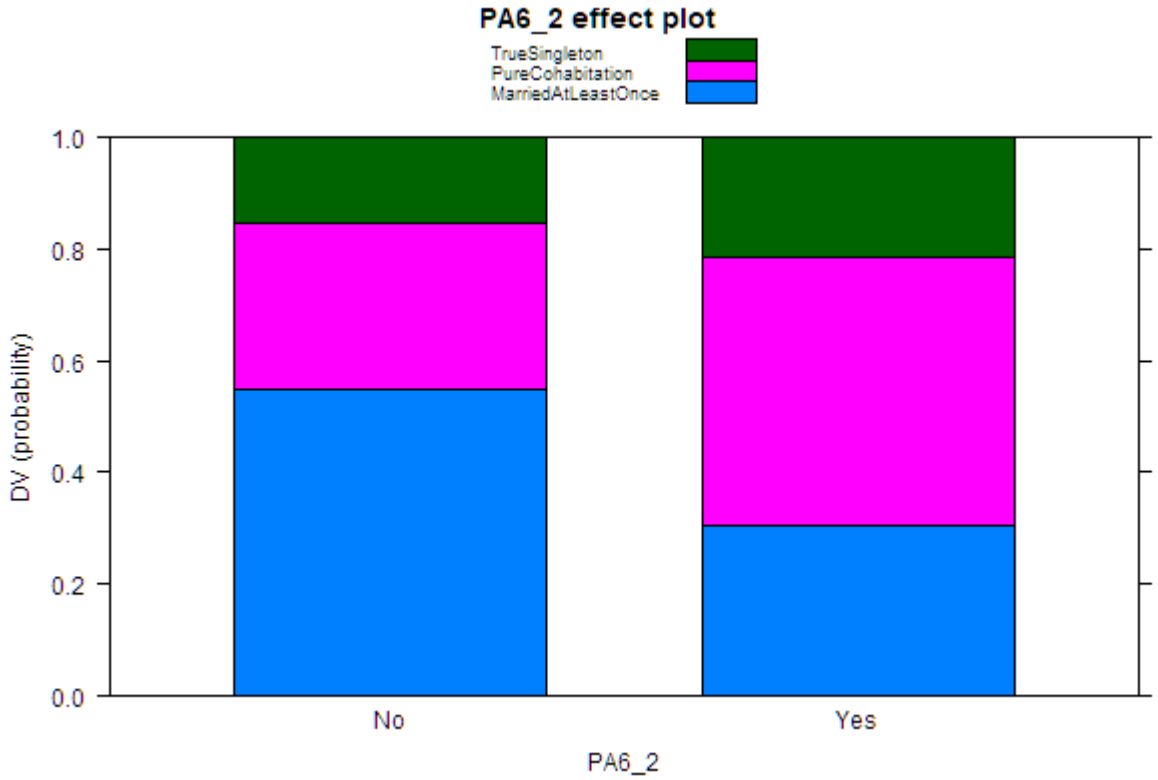
Note. H4SE27 (Wave IV): "In the past 12 months, did you have sex with more than one partner at around the same time?"



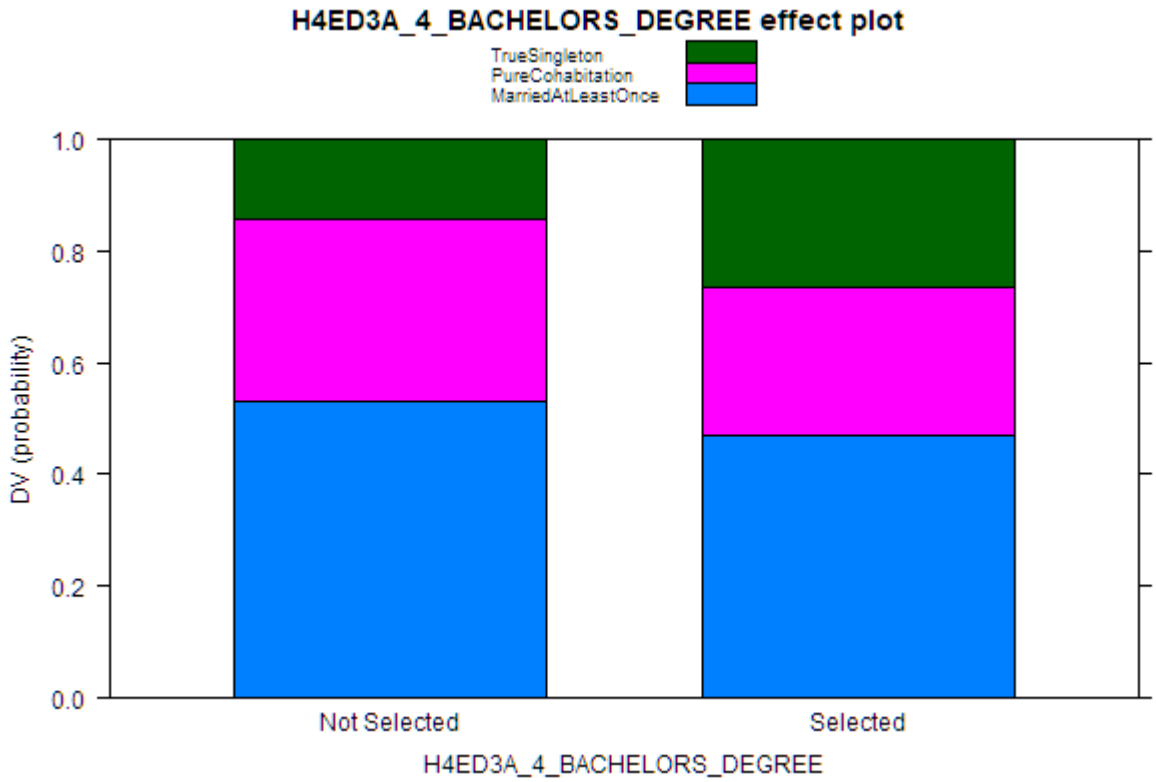
Note. H4IR4 (Wave IV): "Indicate the race of the sample member/respondent from your own observation (not from what the respondent said)." Response: 2 - Black or African American



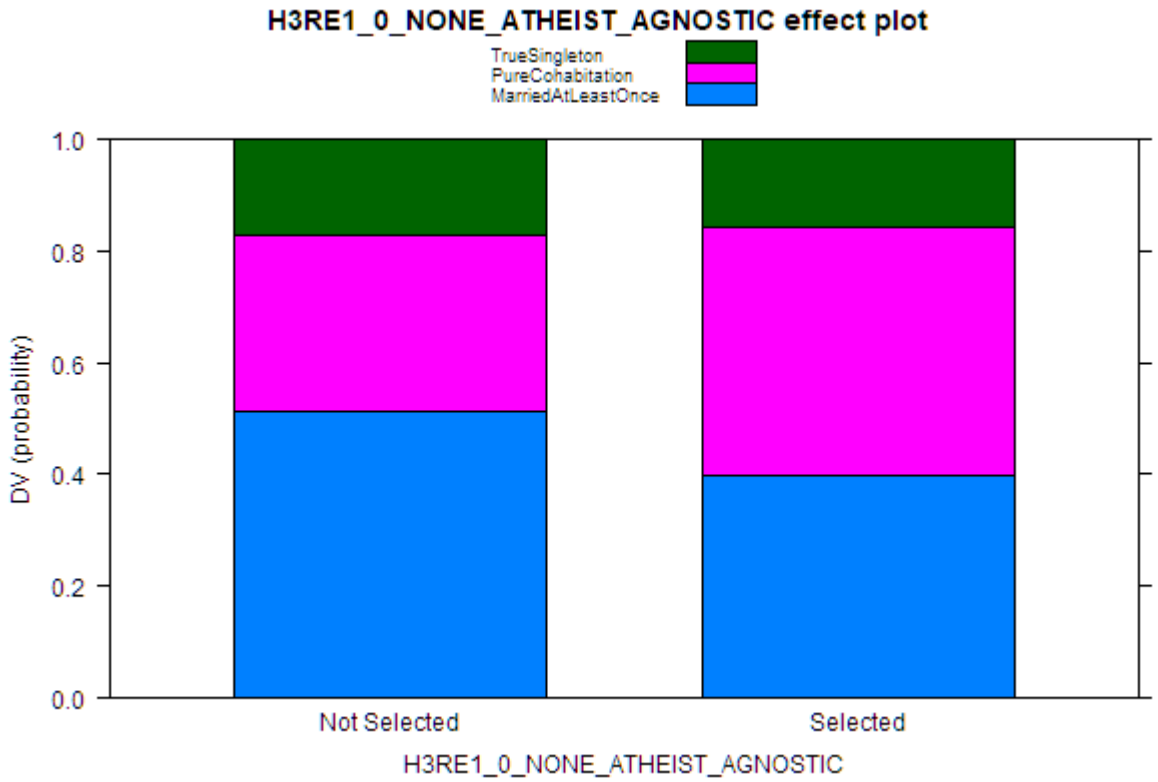
Note. H4SE26V (Wave IV): "In the past 12 months, did you or your partner(s) use any of these methods for birth control or disease prevention (check all that apply): no method used"



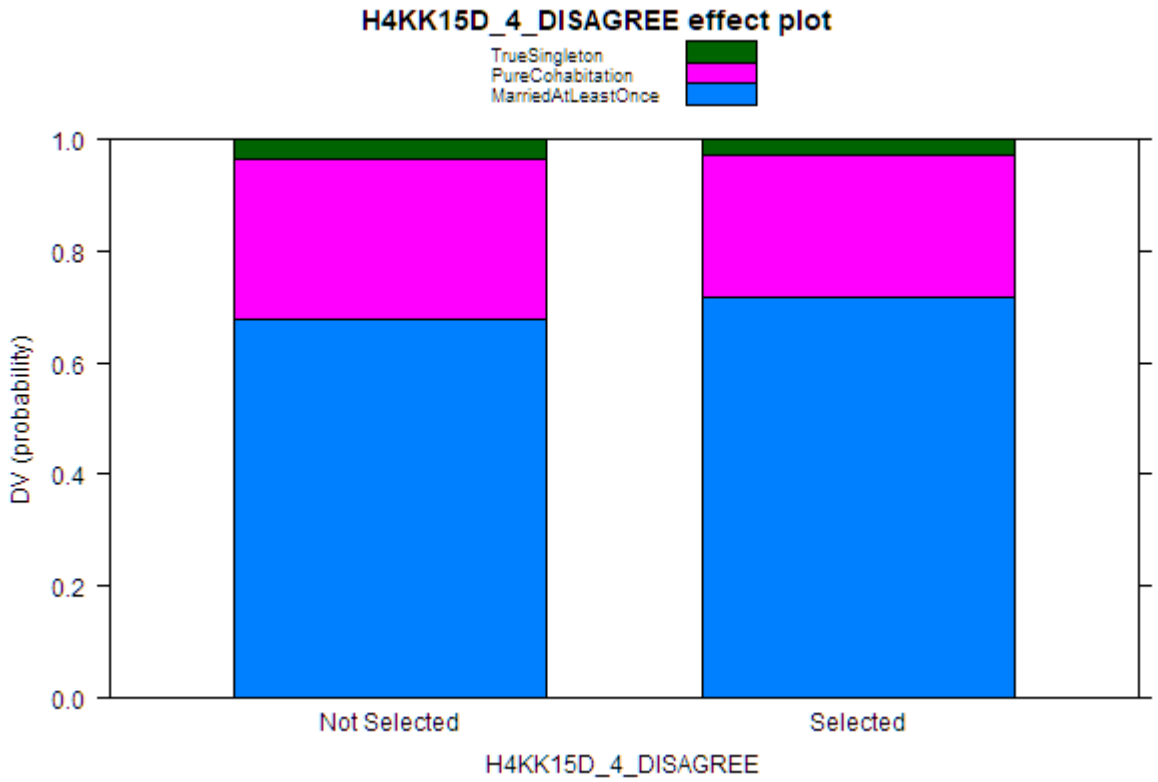
Note. PA6_2 (Wave I): "What is your race (check all that apply): Black/African American"



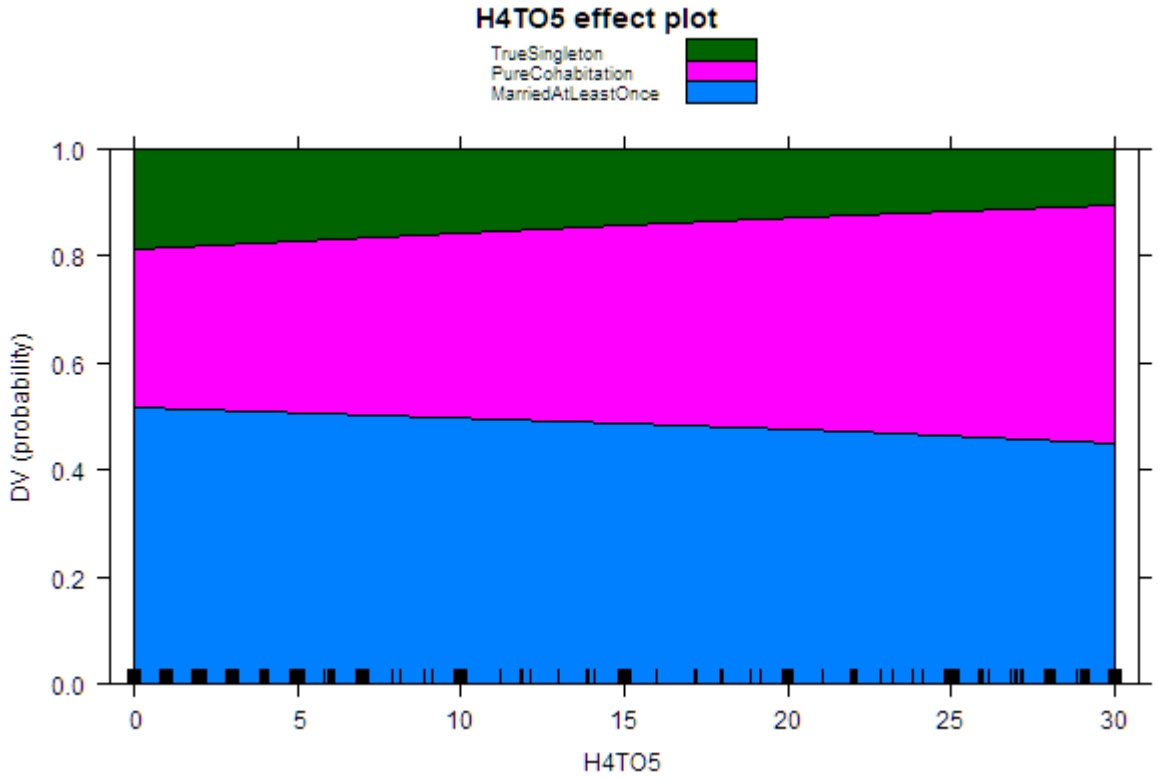
Note. H4ED3A (Wave IV): "Please list all degrees or certificates you have received from a college, university, or vocational/technical school. Do not include certificates you received from programs that lasted less than one year. What is the most recent degree you have received?" Response: 4 - bachelor's degree



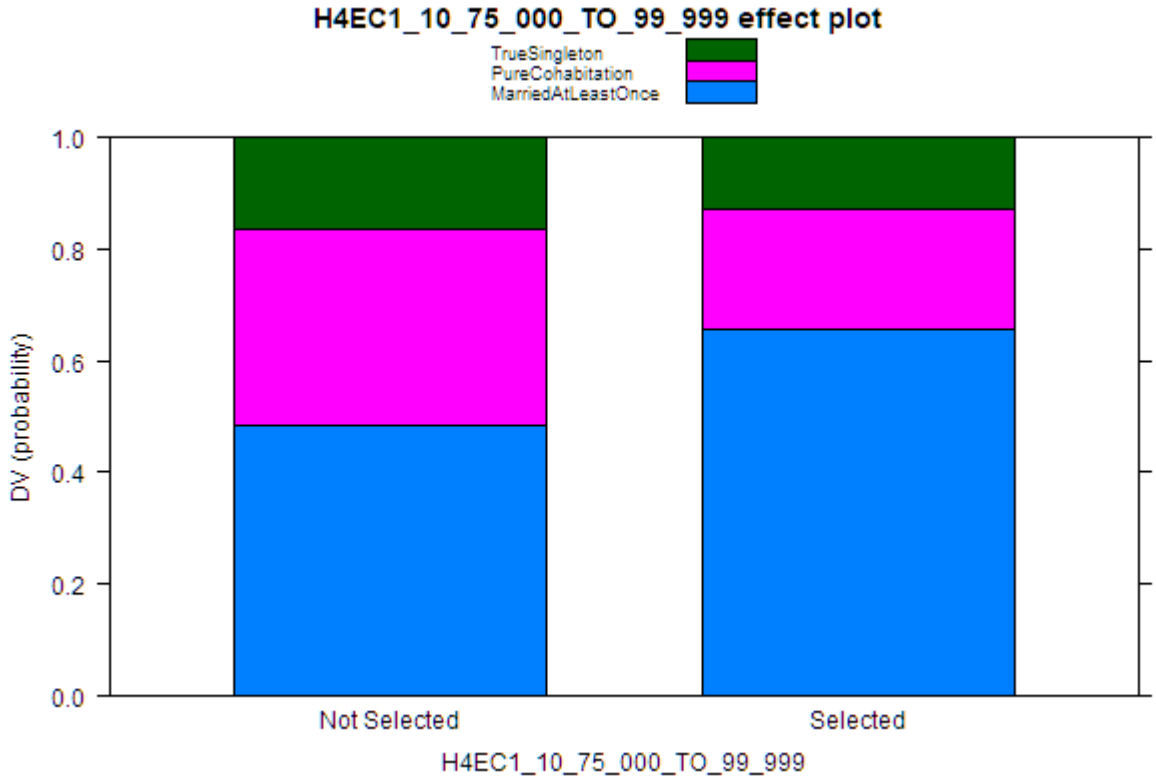
Note. H3RE1 (Wave III): "What is your present religion?" Response: 0 - none, atheist, agnostic



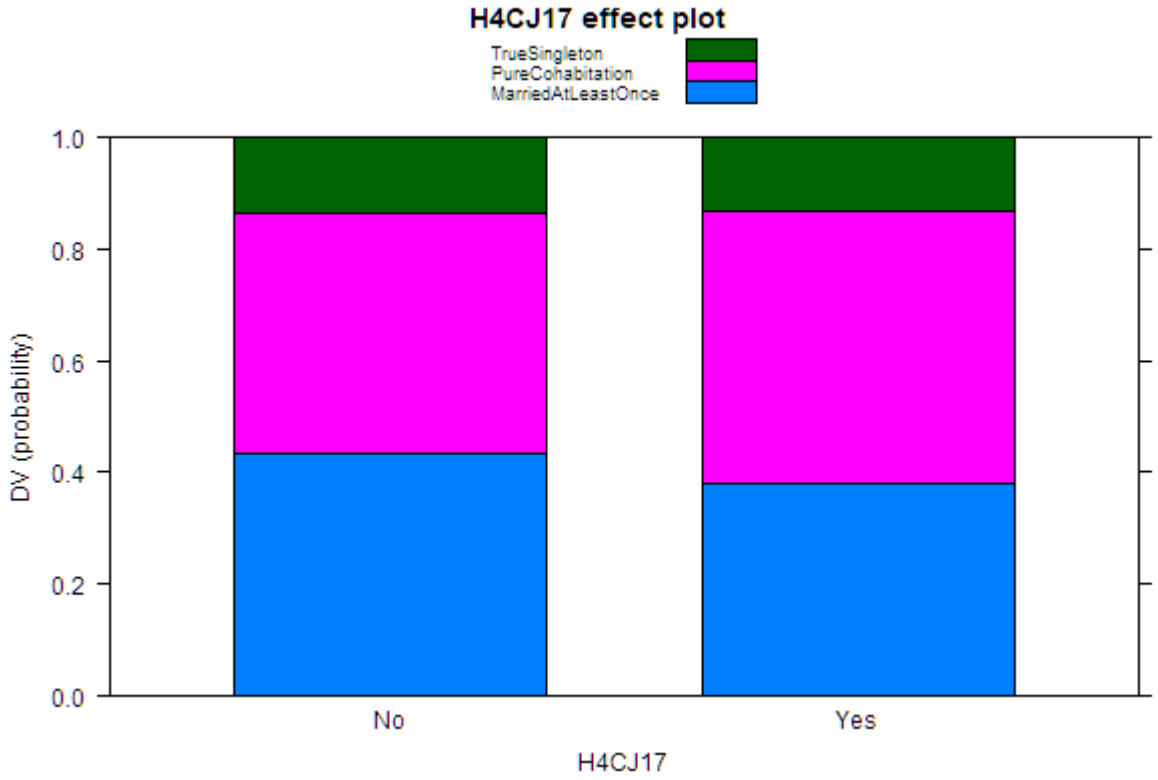
Note. H4KK15D (Wave IV): "How much do you agree or disagree with the following statement? I feel overwhelmed by the responsibility of being a parent." Response: 4 - disagree



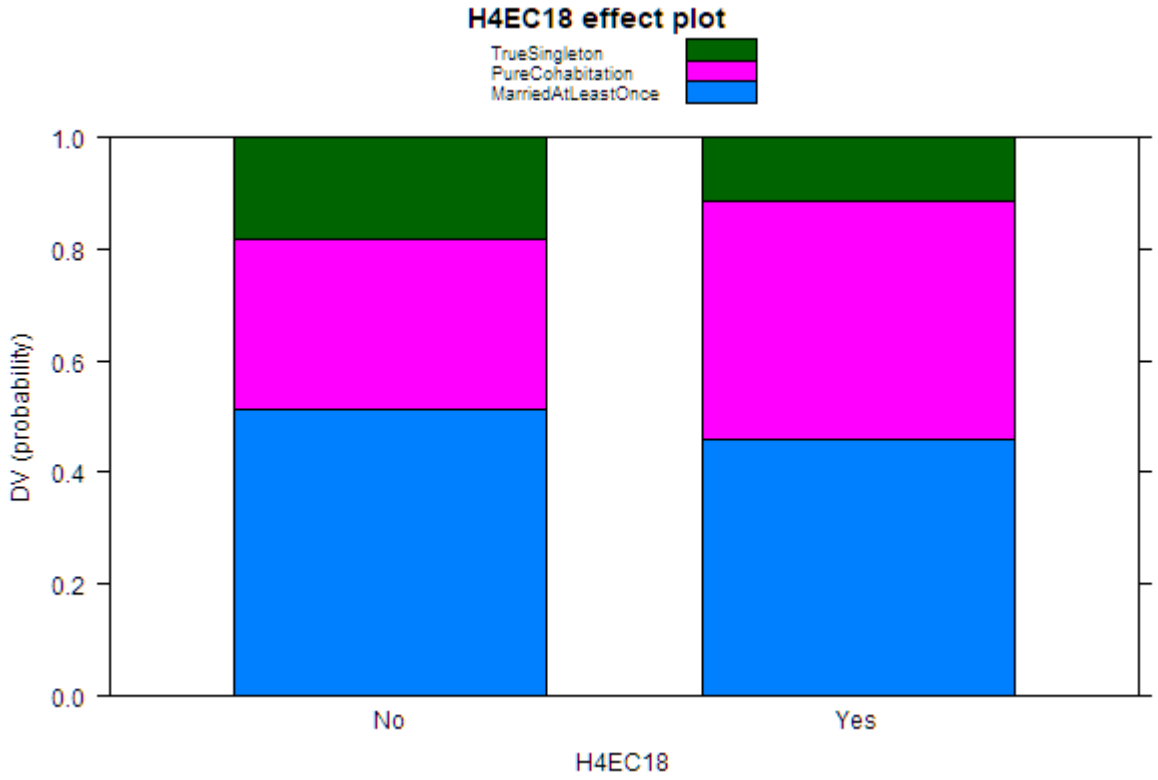
Note. H4TO5 (Wave IV): "During the past 30 days, on how many days did you smoke cigarettes?"



Note. H4EC1 (Wave IV): "Thinking about your income and the income of everyone who lives in your household and contributes to the household budget, what was the total household income before taxes and deductions in {2006/2007/2008}? Include all sources of income, including non-legal sources." Response: 10 - $75,000 to $99,999



Note. H4CJ17 (Wave IV): "Have you ever spent time in a jail, prison, juvenile detention center or other correctional facility?"



Note. H4EC18 (Wave IV): "Between {1995/2002} and {2006/2007/2008}, did you or others in your household receive any public assistance, welfare payments, or food stamps?"