AN EXAMINATION OF POTENTIAL VARIATION IN THE BENEFITS OF HIGHER EDUCATION FOR HEALTH AND WELLBEING

Shawn Bauldry

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the degree of Doctor of Philosophy in the Department of Sociology.

Chapel Hill
2012

Approved by:

Dr. Kenneth A. Bollen

Dr. Michael J. Shanahan

Dr. Guang Guo

Dr. Margarita A. Mooney

Dr. Yang Yang

ABSTRACT

Shawn Bauldry: An Examination of Potential Variation in the Benefits of Higher Education for Health and Wellbeing
(Under the direction of Dr. Kenneth A. Bollen)

Understanding the implications of the significant expansion in higher education over the latter half of the 20th century remains one of the central questions of research in stratification and inequality. Attaining a college degree is associated with numerous advantages ranging from higher earnings to improved health and wellbeing. As higher education continues to expand, however, there is the possibility of increasing variation in the benefits of a college degree. Sociologists have begun to examine variation in the returns to higher education for earnings, civic participation, and fertility. This dissertation contributes to this line of research by analyzing variation in the health-related benefits of a college degree.

Chapters 2 and 3 assess variation in the effects of higher education on health outcomes (self-rated health, systolic blood pressure, body mass index, and smoking) and psychological wellbeing. Data are drawn from the National Longitudinal Study of Adolescent Health. The analyses rely on an innovative approach to detecting variation in the effects of a college degree that is based in the counterfactual framework and uses propensity score models to obtain estimates of various treatment effects.

Chapter 4 examines the potential use of auxiliary variables in the handling of missing data. Missing data is an issue, particularly for the background variables, in the

analyses in Chapters 2 and 3. Methodologists recommend using auxiliary variables, variables that are in some way related to missing data but not otherwise of substantive interest, when addressing missing data. This chapter provides guidelines for when auxiliary variables are more or less likely to be beneficial for reducing bias, which were used to inform the handling of missing data in the earlier chapters.

ACKNOWLEDGEMENTS

Numerous people have provided the intellectual and emotional support and guidance that allowed me to write this dissertation. I owe a debt of gratitude to all of them for helping me find my way through graduate school and learning how to conduct the research outlined in my dissertation.

From the time I arrived at UNC six years ago to begin my graduate work, Ken Bollen has served as my primary advisor. Ken has always been exceptionally generous with his time and has shared his vision of how to conduct research (in both sociology and statistics), how to teach and communicate complex sociological and methodological concepts, and more generally how to be a thoughtful and conscientious scholar in our countless meetings and email exchanges. Ken also encouraged and supported my decision to pursue additional statistical training. Without this investment and Ken's support, I would not have been able to carry out the analyses that underlie Chapters 2 and 3, much less attempt to develop a career as a sociological methodologist. Finally, Ken has generously provided financial support throughout my graduate career.

In the Fall of 2008 I decided to enroll in a workshop devoted to cross-disciplinary work involving genetics and sociology. I had no idea at the time how significant this decision would end up being. It was through this workshop that I began working with Mike Shanahan. We initially shared an interest in understanding some of the genetic underpinnings of educational attainment, but over time our mutual interests

have continued to expand. My interest in health inequalities that underpin this dissertation came out of my work with Mike. Beyond sharing his sociological insights, Mike has been a particularly supportive advisor who managed both to emphasize the progress in my thinking and writing while also encouraging me to keep pushing to produce better work.

Guang Guo, Margarita Mooney, and Yang Yang all provided valuable comments on each of the chapters in this dissertation. Haipeng Shen and Sarah Mustillo also contributed their insights into the work on auxiliary variables, which is reported in Chapter 4. Matt Bradshaw and the members of the stratification and inequality workshop at UNC provided valuable comments on an earlier draft of the work that has become Chapter 2.

A number of fellow graduate students have also had a hand in shaping and supporting this work. We began meeting just to survive our statistics courses, but over time Charles Seguin and Tiantian Yang have had a significant influence on my approach to sociological research. Both Charles and Tiantian always push me to be clearer in my arguments, to keep theory at the forefront, and to think about the bigger picture in my work. I have also benefited from many stimulating conversations about all areas of the discipline with Micah Roos, Brandon Wagner, and Craig Owen.

Navigating graduate school and completing a dissertation is not just about academic relationships. My parents Bill and Sue, fellow academics, have always

encouraged me to pursue my interests, and in many intangible ways have made an academic career appear desirable and possible. Finally, I most want to thank my wife Danielle. Not only has she shared her insight and thoughts that have heavily influenced my work, but she has also provided the daily encouragement and support needed to make it through such a long process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Over the course of the 20th century college enrollment in the U.S. grew from a little over 200,000 people in 1900 to more than 15 million in 2000 (Snyder and Dillow 2011). In percentage terms, about a tenth of a percent of the population aged 15 to 24 were enrolled in college in 1900 as compared with roughly 35 percent of the population aged 15 to 24 in 2000. This expansion in higher education has had profound implications for the life chances of people with and without college degrees (Fischer and Hout 2006). Attaining a college degree is associated with a host of advantages ranging from earning higher income and obtaining more fulfilling jobs to improved psychological wellbeing and health (Blau and Duncan 1967; Hout 1988; Hout and DiPrete 2006; Mirowsky and Ross 1998; Mirowsky and Ross 2003; Preston and Elo 1995).

As higher education continues to expand, some of the people who currently attain degrees would most likely not have in the past. In addition, the recent growth in enrollment has been concentrated among lower tier colleges and universities than in the past. These two considerations raise the possibility that there may be more variation in the benefits of a college degree in recent cohorts than in past cohorts. Estimates of the returns to higher education may be misleading if the potential variation in the returns is related to individual characteristics and attributes.

The primary concern in research on the returns to education, and particularly a college degree, is that the estimates may be upwardly biased due to the presence of unobserved characteristics that influence both attaining a college degree and some favorable outcome, say earnings (Ashenfelter and Krueger 1994; Ashenfelter and Rouse 1998; Griliches 1977; Hauser and Daymount 1977). Another potential source of bias in the estimated returns to higher education that receives less attention arises from the possibility that there are systematic differences in the beneficial effects of a college degree. In general, it is not likely that all people have the same behavioral response to attaining a college degree (Card 1999). Some people will be able to get more from their education than others. As such, standard estimates of the returns to a college degree represent averages over the individual returns in a given population. If the variation in the effects of higher education is not systematic, then the average effect provides a reasonable summary estimate; however, if there is systematic variation, then the average effect will be upwardly biased for some specific subpopulations and downwardly biased for others.

Given the central role of education, and increasingly postsecondary education, in determining one's social position in our society (Blau and Duncan 1967; Fischer and Hout 2006; Hout and DiPrete 2006), as well as the popular emphasis placed on attaining a college degree, it is important to develop a better understanding of the extent of and sources of variation in the returns to a college degree. Sociologists have recently begun to assess general sources of variation in the effects of higher education. In a series of papers, Brand and colleagues find evidence of systematic heterogeneity in the returns to a college degree for income, civic participation, and fertility (Brand 2010;

Brand and Davis 2011; Brand and Xie 2010). Two of the chapters in this dissertation extend this line of research to examine variation in the health-related returns to higher education for four key measures of health behaviors and outcomes (smoking, BMI, self-rated health, and systolic blood pressure) and variation in the protective effects of a college degree for depression.

In addition to determining the potential bias in the estimates of the beneficial effects of a college degree for health and psychological wellbeing, the analytic approached adopted in these chapters allows for an assessment of two key competing theories of potential variation in the health-related returns to higher education – *resource substitution* and *resource multiplication* theory. There are multiple resources that people can draw on to support their health. The distinction between resource substitution theory and resource multiplication theory lies in the relationship between higher education and the other types of resources (e.g., wealth from an advantaged background). If the various resources that can support health serve as substitutes for one another, than having one resource, such as a college degree, can compensate for a lack of other resources, such as wealth obtained from an advantaged background. This is the mechanism that underlies the resource substitution theory. Alternatively, if the various resources that can support health augment each other, then having one resource (e.g., a college degree) can magnify the benefits of other resources (e.g., an advantaged background). This is the mechanism that underlies the resource multiplication theory. Past work testing resource substitution and resource multiplication theory (or similar theories with different terms) have focused on a single attribute (i.e., sex or race) that is linked to differential access to resources (Farmer and

Ferraro 2005; Ross and Mirowsky 2006). The analyses presented in chapters 2 and 3 concerning health behaviors, health outcomes, and depression, provide a more general assessment of these two competing theories (discussed in detail below).

The presence of missing data is a common issue when trying to estimate the returns to higher education, particularly among important covariates such as measures of background socioeconomic status. Contemporary approaches to handling missing data, such as multiple imputation and direct maximum likelihood estimators (Allison 2002; Arbuckle 1996; Little and Rubin 2002; Schafer 1997), allow for the inclusion of variables that are not part of the main analysis, termed "auxiliary variables," in order to improve the performance of the estimators. Methodologists recommend including auxiliary variables to help meet the missing at random (MAR) condition and thereby reduce estimator bias (Allison 2009; Baraldi and Enders 2010; Enders 2010). Relatively little work has been done, however, to establish that conditions under which auxiliary variables can be beneficial in terms of reducing bias and how to assess candidate auxiliary variables in practice. The fourth chapter in this dissertation addresses this gap with new analytic results, supported by a simulation study, that identify when auxiliary variables can reduce bias and recommendations for how to select auxiliary variables in sociological research.

The remainder of the introduction is organized as follows. First, there is a discussion of the links between a college degree and health behaviors, outcomes, and psychological wellbeing. This discussion is followed by a description of the analytic strategy used in the two studies of health and wellbeing. The third section is devoted to the role of auxiliary variables in missing data procedures. Finally, the introduction

concludes with a consideration of the contributions of this work to our understanding of the potential benefits of auxiliary variables in sociological research, health inequalities, and the more general importance of understanding variation in the returns to higher education.

The Benefits of a College Degree for Health and Psychological Wellbeing

A large body of research documents the many links between education and health (Kitagawa and Hauser 1973; Link and Phelan 1995; Mirowsky and Ross 2003; Preston and Elo 1995). Educational attainment is positively associated, for instance, with physical functioning and subjective health, and negatively associated with morbidity, physical impairment, and mortality (Cutler and Lleras-Muney 2008; Elo 2009; Luo and Waite 2005; Mirowsky and Ross 2003; Smith 2007). Numerous studies have also documented a protective effect of education on depression (Bjelland et al. 2008; Kessler et al. 1994; Lorant et al. 2003; Mirowsky and Ross 2003; Turner and Lloyd 1999; Wheaton 1978).

The associations between education and various health outcomes arise, in part, from the beneficial effects of education on health-related behaviors. People with a higher degree are less likely to smoke or to be heavy drinkers, and more likely to have a healthy diet and exercise than people with a high school education (Cutler and Lleras-Muney 2010; Mirowsky and Ross 2003; Pampel, Krueger and Denney 2010). Better educated people are also more likely to engage in health-related preventative care that can have long-lasting benefits (Cutler and Lleras-Muney 2010).

Recent studies indicate that the health-related benefits of education vary systematically across different populations. There has been some debate about how the

effects of education on physical health vary over the life-course and across cohorts (Dupre 2007; Goesling 2007; House, Lantz and Herd 2005; Liu and Hummer 2008; Lynch 2003; Lynch 2006; Mirowsky and Ross 2008; Willson, Shuey and Elder 2007; Zheng and Land 2012). It appears that the effects of education on health increase across cohorts and increase over the life course until a threshold is reached in old age, but more research is needed to make more definitive statements. The benefits of education in limiting depression are also known to vary over the life-course (Bjelland et al. 2008; Miech and Shanahan 2000; Miech et al. 1999).

In addition to variation over the life course and across cohorts, a number of studies find the health-related benefits of education differ across racial and ethnic groups, and males and females. Research has found that blacks do not gain as much from their education as whites, particularly at higher levels of education (Farmer and Ferraro 2005; Liu and Hummer 2008; Shuey and Willson 2008; Williams and Collins 1995). There is also evidence that there is no link between education and depression among blacks, particularly when depression is measured as the experience of major depressive episodes (Cockerham 1990; Hudson et al. 2012; Jackson and Cummings 2011; Williams et al. 2007). Finally, a recent study finds that education provides more a protective effect against depression for females than males (Ross and Mirowsky 2006).

The second and third chapters of the dissertation contribute to our understanding of variation in the health-related benefits of education, specifically a college degree, by adopting an innovative analytic strategy that allows for a more general assessment of variation than has been attempted in past studies. The second

chapter focuses on two important health-related behaviors, smoking and BMI,[1] and two measures of health outcomes, self-rated health and systolic blood pressure. Excess BMI and smoking are two of the leading behavioral causes of death in the U.S. (Cutler and Lleras-Muney 2010; Mokdad et al. 2004; Pampel, Krueger and Denney 2010). Self-rated health has proven to be a reliable leading indicator of morbidity and mortality (Ferraro and Farmer 1999; Idler and Benyamini 1997; Jylhä 2009; Jylhä, Volpato and Guralnik 2006). Systolic blood pressure provides a more objective measure of health that is associated with cardiovascular disease and strokes (Fields et al. 2004; Lawes, Vander Hoorn and Rodgers 2008).

The third chapter focuses on variation in the protective effects of a college degree against depression. Depression is one of the leading causes of disability in the contemporary US, with close to 1 in 10 adults suffering from depressive symptoms (Strine et al. 2008; World Health Organization 2004). The analysis of depression relies on the same analytic strategy as the analysis of health behaviors and outcomes, but in addition addresses measurement error in the indicators of depressive symptoms (items from the Center for Epidemiologic Studies Depression scale).

Identifying Variation in the Health-Related Returns to Higher Education

The analyses in the second and third chapters of the dissertation rely on an innovative analytic approach for detecting consequential variation, or heterogeneity, in the health-related returns to a college degree. The approach is rooted in the counterfactual framework for analysis (Heckman 2005; Morgan and Todd 2008;

[1] BMI is, of course, not a behavior, but rather reflects behavior related to diet and physical activity.

Morgan and Winship 2007; Rubin 1974; Splawa-Neyman [1923] 1990). Within the counterfactual framework, attaining a college degree can be thought of as a "treatment" – i.e., something that in principle can be manipulated (Holland 1986). The analogue to the standard regression-based estimate of the returns to higher education is the average treatment effect (ATE). As implied by the label, the ATE is an estimate of the average effect of a college degree on some outcome over a given population. It is, however, possible to define other treatment effects that average over different subpopulations.

The analytic approach in the analyses of the health-related benefits of a college degree rests on assessing the difference between two alternative treatment effects: the average treatment effect for the treated (ATT) and the average treatment effect for the controls (ATC) (Morgan and Todd 2008). In these analyses, the ATT provides an estimate of how much a randomly chosen person who attained a college degree benefits from that degree with respect to a given health outcome. Similarly, the ATC provides an estimate of how much a randomly chosen person who did not attain a college degree would benefit from that degree with respect to a given health outcome if s/he were to attain a degree. A difference in the ATT and ATC provides evidence that the overall estimate of the health-related returns to a college degree, the average treatment effect (ATE), masks important dimensions of heterogeneity and provides a misleading or incomplete sense of the benefits of a college degree.

In addition to identifying variation in the ATE, a comparison of the ATT and ATC also allows one to assess the relative importance of resource substitution theory and resource multiplication theory. The *resource substitution* theory holds that a college

degree can serve to compensate for other health-related disadvantages (Mirowsky and Ross 2003; Ross and Mirowsky 2006; Ross and Mirowsky 2010). In contrast, the *resource multiplication* theory holds that a college degree simply magnifies other health-related advantages (Mirowsky and Ross 2003; Ross and Mirowsky 2006; Ross and Mirowsky 2010). An estimated ATT that is greater than the ATC implies that people who have the most to gain from a college degree are the ones attaining degrees. This is typically termed "positive selection" and is consistent with the resource multiplication hypothesis. Alternatively, an estimated ATC that is greater than the ATT implies that people who stand to gain the most from a college degree are not the ones who actually attain a degree. This is typically termed "negative selection" and is consistent with the resource substitution hypothesis.

Estimates of the ATE, ATT, and ATC are obtained through the use of weights constructed from propensity score models (Guo and Fraser 2010; Morgan and Winship 2007; Rosenbaum and Rubin 1983; Smith 1997). For this analysis, propensity score models are simply logit models that estimate the probability of attaining a college degree. The use of propensity score models allows for a determination of how well the covariates adjust for preexisting differences between people who do and do not attain a college degree. The propensity score models in these analyses are not, however, used to advance claims that the effects are causal any more so than would be reasonable with standard regression models. Rather the propensity score approach serves to highlight the assumptions that underlie treating the estimates as causal effects.

Auxiliary Variables in the Handling of Missing Data

As with much quantitative sociological work, missing data is an issue in the analyses of health behaviors, health outcomes, and depression. In particular, the propensity score models rely on a rich set of covariates to capture the primary predictors of attaining a college that are also related to the health measures. Most of the covariates are not missing for a high percentage of respondents (i.e., less than 3-5 percent missing), but a few, particularly variables measuring the socioeconomic status of the family in which the respondent grew up are missing for substantially more respondents (i.e., 20-25 percent missing). This extent of missing data should not be ignored.

Social scientists increasingly rely on multiple imputation (MI) to address missing data in their analyses. MI provides estimators with desirable properties under less restrictive assumptions than other approaches (e.g., complete-case analysis or pairwise deletion) to handling missing data (Allison 2002; Little and Rubin 2002). Provided certain assumptions about missingness are met, MI produces estimators that are consistent, asymptotically efficient, and asymptotically normal as the number of imputations approaches infinity (Allison 2002; Little and Rubin 2002; Schafer 1997). With respect to missing data, MI estimators obtain these properties only under the assumption that the data are missing at random (MAR) or missing completely at random (MCAR) (Rubin 1976).

Data are MCAR if the probability of missingness is independent of the observed variables and independent of the values that are missing. Data are MAR if the probability of missingness is independent of the values that are missing conditional on

the observed values. MAR is a less stringent criterion than MCAR, but it is not possible to test empirically whether data meet the MAR. The inability to test whether the MAR condition is met means that analysts must strive to include causes or correlates of missing values in their models. If an important cause of missingness is left out of the model, then the estimator can be badly biased and the resulting estimates can be misleading.

In some cases, however, variables related to missingness may be not of substantive or theoretical interest. Methodologists refer to such variables as auxiliary variables (AVs) and recommend that analysts identify and use AVs when addressing missing data (Collins, Schafer and Kam 2001; King et al. 2001; Rubin 1996; Schafer and Graham 2002; Schafer and Olsen 1998). In their seminal study, Collins et al. (2001) defined two types of AVs. Type 1 AVs consist of variables associated with both a variable missing data and the pattern of missingness in that variable. Type 2 AVs consist of variables associated with a variable missing data, but not with the pattern of missingness. They found that under certain conditions concerning the pattern of missingness in the data, the extent of missing data, and the strength of the associations with the AV, Type 1 AVs can have an appreciable effect on reducing bias in an estimator and both Type 1 and Type 2 AVs can improve the efficiency of an estimator. Some recent studies, however, have questioned the usefulness of AVs due to the stringent nature of the conditions in which they have been shown to reduce bias and the typically small efficiency gains. In simulation studies with arguably more realistic conditions for applied work, the benefits of AVs appear to be minimal (Mustillo forthcoming; Savalei and Bentler 2009; von Hippel 2007).

The fourth chapter of the dissertation builds on prior studies in three ways to inform the handling of missing data in the second and third chapters and more generally to provide guidelines for sociologists looking to incorporate AVs into their handling of missing data with the specific goal of potentially reducing bias in an estimator. First, the chapter offers a systematic examination of the properties an AV should satisfy in order to help an analyst meet the MAR condition and therefore reduce bias. This examination reveals that the properties are more strict than has been appreciated in past simulation studies and that some of the claims from the previous studies are inaccurate. Second, the chapter presents the results from a simulation study that isolates and separately manipulates the associations between the AV and the variable missing data and between the AV and missingness. Past simulation studies have not isolated these two associations and therefore obscured potentially important dimensions in the performance of AVs. Third, the chapter provides an empirical example involving AVs that illustrates a procedure for assessing candidate AVs.

Contributions

This dissertation contributes to our understanding of health inequalities through an examination of variation in the health-related returns to a college degree in two ways. First, the presence of systematic variation in the beneficial effects of higher education for health behaviors, health outcomes, and depression serves as an indication that past estimates of the effects of higher education are biased for some subpopulations. Second, the analytic strategy used to identify systematic variation in the health-related returns to a college degree also allows for an assessment of whether higher education serves to alleviate or exacerbate preexisting health inequalities.

More generally, the focus in this dissertation on variation in the effects of a college degree highlights one of the potential ramifications of the continuing expansion of higher education. Education, and particularly higher education, plays a central role in the stratification system in the U.S. (and increasingly throughout the world). Sociologists have recently begun to examine the extent of variation in the returns to education across a variety of outcomes. This dissertation contributes to that general program of developing a deeper understanding of the role of higher education in the contemporary U.S.

This dissertation also makes a methodological contribution with regard to the conditions under which auxiliary variables can be beneficial when addressing missing data in sociological analyses. The two most common contemporary approaches to handing missing data, multiple imputation (MI) and direct maximum likelihood (DML), rest on the assumption that the data are missing at random. In addition, it is possible to incorporate variables that are not part of the primary analysis in MI and DML estimators, which allows for the use of auxiliary variables to improve the performance of these estimators. As sociologists increasingly rely on MI and DML to address missing data, a more complete understanding of when auxiliary variables may be beneficial is important.

Organization of the Dissertation

The next chapter turns to the analysis of variation in the returns to a college degree for health behaviors and outcomes. The third chapter reports on the analysis of variation in the protective effects of higher education against depression. Chapter four presents the analytic results, simulation results, and empirical example related to the

use of auxiliary variables for handling missing data in sociological research. Finally, chapter 5 offers a general conclusion for the dissertation that brings together the key findings from each of the chapter and discusses the more general contributions of this research.

CHAPTER II: HEALTH-RELATED RETURNS TO A COLLEGE DEGREE

A large body of research documents the many links between education and health (Kitagawa and Hauser 1973; Link and Phelan 1995; Mirowsky and Ross 2003; Preston and Elo 1995). Educational attainment is positively associated, for instance, with physical functioning and subjective health, and negatively associated with morbidity, physical impairment, and mortality (Cutler and Lleras-Muney 2008; Elo 2009; Luo and Waite 2005; Mirowsky and Ross 2003; Smith 2007). These associations arise, in part, from the beneficial effects of education on health behaviors. People with a higher degree are less likely to smoke or to be heavy drinkers, and more likely to have a healthy diet and exercise than people with a high school education (Cutler and Lleras-Muney 2010; Mirowsky and Ross 2003; Pampel, Krueger and Denney 2010). Better educated people are also more likely to engage in health-related preventative care that can have long-lasting benefits (Cutler and Lleras-Muney 2010).

Recent studies indicate that the benefits of education for health and health behaviors vary systematically across different populations. There has been some debate about how the effects of education on health vary over the life-course and across cohorts (Goesling 2007; House, Lantz and Herd 2005; Liu and Hummer 2008; Lynch 2003; Lynch 2006; Mirowsky and Ross 2008; Willson, Shuey and Elder 2007; Zheng and Land 2012). It appears that the effects of education on health increase across

cohorts and increase over the life course until a threshold is reached in old age, but more research is needed to make more definitive statements. In addition to variation over the life course and across cohorts, a number of studies find the health-related benefits of education differ across racial and ethnic groups. Research has found that blacks do not gain as much from their education as whites, particularly at higher levels of education (Farmer and Ferraro 2005; Liu and Hummer 2008; Shuey and Willson 2008; Williams and Collins 1995).

This paper makes two contributions to our understanding of the health-related returns to education and variation in the returns. First, the paper adopts an innovative approach to assessing the presence of variation in the benefits of a college degree for health and health behaviors (Morgan and Todd 2008). The analysis centers on attaining a four-year college degree for two reasons. In recent cohorts, a college degree represents a key educational threshold that differentiates the life chances of individuals (Fischer and Hout 2006). The expansion in higher education has led to people attaining degrees who would not have in the past, which suggests that there may be more variation in general in the effects of a college degree among recent cohorts than one would expect among lower levels of education.

The approach to assessing heterogeneity in the effects of a college degree is rooted in the counterfactual framework for analysis (Heckman 2005; Morgan and Winship 2007; Rubin 1974; Splawa-Neyman [1923] 1990) and relies on comparing estimates of the average treatment effect for the treated (ATT) with the average treatment effect for the controls (ATC). In this analysis, the ATT provides an estimate of how much a randomly chosen person with a college degree (the treatment) benefits

from that degree with respect to a given health behavior or health outcome. Similarly, the ATC provides an estimate of how much a randomly chosen person without a college degree would benefit from that degree with respect to a given health behavior or health outcome. A difference in the ATT and ATC provides evidence of important dimensions of variation in the health-related returns to education (discussed in detail below).

Second, a comparison of the ATT and ATC also provides a test of the relative significance of two competing theories concerning the role of a college degree in alleviating or exacerbating health inequalities. The *resource substitution* theory holds that a college degree can serve to compensate for other health-related disadvantages (Mirowsky and Ross 2003; Ross and Mirowsky 2006; Ross and Mirowsky 2010). In contrast, the *resource multiplication* theory holds that a college degree simply magnifies other health-related advantages (Mirowsky and Ross 2003; Ross and Mirowsky 2006; Ross and Mirowsky 2010).[2] An estimated ATT that is greater than the ATC implies that people who have the most to gain from a college degree are the ones attaining degrees. This is typically termed "positive selection" and is consistent with the resource multiplication hypothesis. Alternatively, an estimated ATC that is greater than the ATT implies that people who stand to gain the most from a college degree are not the ones who actually attain a degree. This is typically termed "negative selection" and is consistent with the resource substitution hypothesis. As such, this paper provides new evidence for the relative importance of the resource substitution and multiplication hypotheses.

[2] This is quite similar to the *diminishing returns* hypothesis outlined by Farmer and Ferraro (2005).

Data for this study come from Waves 1 and 4 of the National Longitudinal Study of Adolescent Health (Add Health), a nationally representative sample of young adults. Wave 4 was collected in 2007 and 2008 when respondents were in their late 20s and early 30s. This represents a recent cohort and an age when most respondents are old enough to have completed their education. The analysis examines two measures of health behaviors, body mass index (BMI)[3] and smoking, and two measures of health outcomes, self-rated health and systolic blood pressure. Excess BMI and smoking are two of the leading behavioral causes of death in the U.S. (Cutler and Lleras-Muney 2010; Mokdad et al. 2004; Pampel, Krueger and Denney 2010). Self-rated health has proven to be a reliable leading indicator of morbidity and mortality (Ferraro and Farmer 1999; Idler and Benyamini 1997; Jylhä 2009; Jylhä, Volpato and Guralnik 2006). Systolic blood pressure provides a more objective measure of health that is associated with cardiovascular disease. Although Add Health respondents are young to experience problems with systolic blood pressure, recent analyses of Add Health respondents have found elevated levels of blood pressure and a higher prevalence of hypertension for their age group than in past cohorts (Nguyen et al. 2011). Given the known differences in these health behaviors and outcomes and variation in the effects of education by sex and race/ethnicity, all analyses are conducted separately for black and white, men and women.

[3] BMI is, of course, not a behavior, but rather reflects behaviors related to diet and exercise.

Mechanisms Linking Higher Education to Health and Health Behaviors

Health researchers have proposed and tested a number of mechanisms that give rise to the health-related benefits of a higher degree. The mechanisms can be broadly grouped into two categories: (1) economic resources and (2) social psychological and relational resources (Mirowsky and Ross 2003; Ross and Mirowsky 2010). Higher education, particularly in recent cohorts, plays a central role in providing access to high quality jobs with a decent income and a potentially creative and autonomous work environment. The economic resources that accrue from such positions provide a buffer from the stress associated with poverty and economic hardship that can lead to a variety of negative health behaviors and health outcomes (Lantz et al. 2005; Mirowsky and Ross 2003; Pampel, Krueger and Denney 2010). In addition, economic resources may directly lead to better health behaviors and outcomes by facilitating healthy behaviors (e.g., membership to a gym) and allowing greater access to preventative medical care (Cutler and Lleras-Muney 2008; Freese and Lutfey 2011; Mirowsky and Ross 2003; Pampel, Krueger and Denney 2010). A recent study estimates that economic resources, in particular income, access to health insurance, and resources from one's family background, account for about thirty percent of the relationship between education and health-related behaviors (Cutler and Lleras-Muney 2010).

A college degree may also promote better health behaviors and health outcomes through the development of social psychological and relational resources. Higher education has the potential to instill a sense of mastery and personal control that give people the motivation, knowledge, and ability to address their health-related needs (Cutler and Lleras-Muney 2008; Mirowsky and Ross 1998; Mirowsky and Ross 2003;

Pampel, Krueger and Denney 2010; Ross and Mirowsky 2010). While pursing a college education many people form relationships with other people in college and often maintain relationships with other highly educated people after completing a degree. This leads people with higher degrees to have social networks with other health-oriented people that can support healthy behaviors and develop group-based norms that protect against unhealthy behaviors (Freese and Lutfey 2011; Kawachi, Subramanian and Kim 2008; Smith and Christakis 2008). For instance, several studies have identified network effects on smoking habits and obesity (Boardman et al. 2005; Christakis and Fowler 2007; Christakis and Fowler 2008; Cutler and Glaeser 2007). More generally, social support can help alleviate stress and appears to reduce mortality, though the links between social support and physical health are more contentious (House, Landis and Umberson 1988; Seeman 1996).

Variation in the Health-Related Returns to a College Degree

The dramatic increase in college enrollment over the course of the 20th century has led to changes in who attains college degrees and the nature of colleges and universities granting degrees (Snyder and Dillow 2011). In percentage terms, about a tenth of a percent of the population aged 15 to 24 were enrolled in college in 1900 as compared with roughly 35 percent of the population aged 15 to 24 in 2000. Some people who currently attain degrees would most likely not have in the past. In addition, the recent expansion in higher education has been concentrated among lower tier colleges and universities. Taken together, these changes raise the possibility that there may be more variation in the benefits of a college degree in recent cohorts than in past cohorts.

Changes in the composition of people attaining degrees and in the institutions granting degrees could affect the health-related returns to a college degree through both economic and social psychological and relational resources. For some people, the economic returns to a college degree may not be as high as for others, in which case a college degree may not provide as much protection from economic hardship and resources to afford a healthy lifestyle. In addition, a college degree may not provide the same sense of mastery and personal control, nor does it necessarily provide access to health-promoting social networks among recent cohorts. Thus, an increase in the variation of the benefits of a college degree among recent cohorts has an impact through a variety of mechanisms that link education to health-related behaviors and health outcomes.

Resource Substitution and Resource Multiplication Theories

In addition to assessing the extent of variation in the health-related benefits of a college degree, the analytic approach used in this paper also allows for an assessment of the relative strength of two competing theories of the role of education in alleviating or exacerbating health inequalities: the *resource substitution* and the *resource multiplication* theories (Mirowsky and Ross 2003; Ross and Mirowsky 1989; Ross and Mirowsky 2010). There are potentially multiple resources that people can draw on to support their health. The distinction between resource substitution and resource multiplication theories lies in the relationship between higher education and the other types of resources (e.g., wealth from an advantaged background). If the various resources that can support health serve as substitutes for one another, than having one resource, such as a college degree, can compensate for a lack of other resources, such as

wealth obtained from an advantaged background. This is the mechanism that underlies the resource substitution theory. Alternatively, if the various resources that can support health augment each other, then having one resource (e.g., a college degree) can magnify the benefits of other resources (e.g., an advantaged background). This is the mechanism that underlies the resource multiplication theory.

A comparison of the estimates of the health-related returns of a college degree for people who attained a degree (the ATT) with the health-related returns of a college degree for people who did not attain a degree (the ATC) not only provides evidence of variation in the average effect of a college degree but also allows for an assessment of resource substitution and multiplication theories. If the ATC is greater than the ATT, then this indicates that a college degree would provide greater health-related returns to people who did not attain a degree were they to do so than for people who did attain a degree. This pattern is consistent with the resource substitution theory because people who do not attain a college degree have fewer resources on average than people who do attain a college degree. On the other hand, if the ATT is greater than the ATC, then this indicates that a college degree provides greater health-related returns to people who attain a degree than would be realized by people who did not attain a degree. This pattern is consistent with the resource multiplication hypothesis because it suggests that a college degree amplifies pre-existing advantages (other resources).

Methodological Issues

The standard approach to analyzing the effect of a college degree on a particular health outcome or behavior is to estimate a regression model of the form

$$y_i = \alpha + \gamma D_i + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \qquad (2.1)$$

where y_i is a health outcome of interest,[4] D_i is an indicator for attaining a college degree, $\mathbf{x}_i$ is a vector of covariates that are associated with both the health outcome and the probability of attaining a college degree, and ε_i is a disturbance term. In this equation, if one assumes that there are no omitted variables (i.e., $\text{Cov}(D_i, \varepsilon_i) = 0$ and $\text{Cov}(\mathbf{x}_i, \varepsilon_i) = 0$), then $\hat{\gamma}$ provides an estimate of the average causal effect of attaining a college degree on a given health outcome.

The standard approach has a straightforward interpretation in the counterfactual framework (Heckman 2005; Morgan and Winship 2007; Rubin 1974; Splawa-Neyman [1923] 1990). Attaining a college degree can be thought of as a "treatment" – i.e., something that in principle can be manipulated (Holland 1986). The health benefits that arise from receiving the treatment, attaining a college degree, may vary from person to person. It is not, in general, possible or desirable to estimate a separate causal effect from a treatment for all individuals. Instead, analysts are typically interested in an average causal effect defined over a given population or subpopulation (Heckman 2005). This can be reflected in the model in equation (1) by simply adding an index j to γ that references the population or subpopulation of interest.

Identifying Variation in the Health-Related Returns to a College Degree

People who receive the treatment and people who do not receive the treatment are two common subpopulations of interest in the counterfactual framework. In this analysis the treatment is the attainment of a college degree, so the two subpopulations

[4] If the health outcome is not a continuous variable (e.g., an indicator for hypertension) or represents the time to an event (e.g., time to death), then (1) should be understood to refer to an appropriate regression model (e.g., a logistic regression model or a Cox regression model).

are people who did and did not attain a degree. The average causal effect for people who attained a college degree, the "average treatment effect for the treated" (ATT), is an estimate of how much a randomly selected person who attained a college degree benefited from that degree with respect to their health (Heckman 2005; Morgan and Harding 2006). Similarly, the average causal effect for people who did not attain a college degree, the "average treatment effect for the controls" (ATC), is an estimate how much a randomly selected person who did not attain a college degree would benefit if she were to attain one (Heckman 2005; Morgan and Harding 2006). A comparison of the ATT and ATC provides a means of assessing whether there is consequential variation in the average health-related returns to a college degree.

There are two potential sources of heterogeneity in the effect of a college degree on health that could lead to different estimates for the ATT and ATC (Morgan and Todd 2008; Morgan and Winship 2007). First, a difference in the ATT and ATC could arise if the effect of a college degree depends on one of the other covariates in the model. In other words, there is an interaction effect between the treatment, attaining a college degree, and at least one of the other predictors of the health outcome. Second, a difference in the ATT and ATC could arise if the effect of a college degree depends on an unobserved variable that predicts attaining a college degree. Methodologists divide causal effect heterogeneity due to unobserved variables into two types: (1) heterogeneity due to an omitted variable and (2) heterogeneity due to the possibility that individuals anticipate the potential gains from the treatment (or control) and select into the treatment (or control) to realize these gains (Heckman 2005; Heckman, Urzua and Vytlacil 2006; Morgan and Todd 2008; Morgan and Winship 2007). A comparison

of the ATT and the ATC allows for a more complete assessment of potential variation in the protective effects of college degree and avoids the "curse of dimensionality" that comes with attempting to include all plausible interactions with observed variables.

Data and Methods

Data

This analysis draws on Waves 1 and 4 from the National Longitudinal Study of Adolescent Health (Add Health).[5] The first wave of Add Health was based on a nationally representative sample of youth in grades 7 through 12 in the United States in 1994. The fourth wave of data collection occurred between January 2008 and February 2009, roughly 14 years after the first wave. About 75 percent of the 20,745 youth interviewed at Wave 1 were re-interviewed at Wave 4, resulting in 15,701 respondents with data from both waves.

Past research has found that the effects of education on health behaviors and health outcomes differ by sex and race (Farmer and Ferraro 2005; Ross and Mirowsky 2006; Shuey and Willson 2008; Williams and Collins 1995). Because of this, the analysis sample was restricted to white and black, males and females (N = 11,529) and separate models were estimated for each of the four subgroups. In addition, Add Health provides sample weights to adjust for the differing probabilities of inclusion in the original sample frame and attrition between Waves 1 and 4. Respondents (N = 674) who did not

[5] See Harris, Halpern, Whitsel, Hussey, Tabor, Entzel, and Udry Harris, Kathleen Mullan, Carolyn T. Halpern, Eric A. Whitsel, Jon M. Hussey, Joyce W. Tabor, Pamela P. Entzel, and J. Richard Udry. 2009. "The National Longitudinal Study of Adolescent Health: Research Design." URL: http://www.cpc.unc.edu/projects/addhealth/design. for a detailed description of the construction of the Add Health sample.

have sample weights were excluded from the analysis sample. Finally, 2 respondents missing information about the highest educational degree completed were excluded from the analysis sample.[6] This results in an analysis sample of N = 10,853.

All of the remaining missing data were addressed using multiple imputation (Little and Rubin 2002). Ten complete data sets were constructed using the chained equation approach implemented in Stata 12 (StataCorp 2011). Most of the measures used in the analysis were missing for less than 4 percent of the cases. The exceptions were some of the measures related to parental SES and the respondent's birth weight. These variables were missing for 20 to 25 percent of the cases.[7] Multiple imputation appeared to perform reasonably well for these variables. With the exception of a small number of outliers (i.e., less than 10), the range of the imputed values was quite similar to the range of the non-imputed values for all of these variables. In addition, the mean and variance of the measures for the non-imputed cases were quite similar (i.e., within a couple hundredths of a point) to the mean and variance of the measures with the imputed cases included.

Variables

Health Measures. This analysis focuses on two measures of health and two measures of health behaviors. The first measure, self-rated health, is a general measure

[6] This missing data could not be handled with multiple imputation because the estimation of the propensity score weights and subsequent models does not allow for potential differences in educational level across for the same case across imputed data sets.

[7] Auxiliary variables that could improve the handling of missing data were considered, but none met the required associations outlined in chapter 4 to have an impact on reducing bias.

of health based on responses to the question "In general, how is your health?" that include "excellent" to "poor." Self-rated health has proven to be a reliable leading indicator of morbidity and mortality (Ferraro and Farmer 1999; Idler and Benyamini 1997; Jylhä 2009; Jylhä, Volpato and Guralnik 2006). Self-rated health is treated as a dichotomous variable ("excellent" or "very good" versus "good," "fair," or "poor"), which is consistent with the majority of past research using this item.[8] The second measure is systolic blood pressure (SBP),[9] a more objective measure of health associated with cardiovascular disease. Recent analyses of Add Health respondents have found elevated levels of blood pressure and a higher prevalence of hypertension for their age group than in past cohorts (Nguyen et al. 2011).

The analysis focuses on two measures of health behaviors that are among the leading behavioral causes of death in the U.S. (Cutler and Lleras-Muney 2010; Mokdad et al. 2004; Pampel, Krueger and Denney 2010). The first, body mass index (BMI), is a commonly used measure that reflects behaviors related to diet and exercise. The second measure is smoking. Smoking is coded as an indicator that takes a value of 1 for respondents who reported smoking at least once a day over the past month.

College Education. In Wave 4 respondents were asked the highest educational level of education they had achieved to date. This information was used to construct an

[8] Supplemental analyses were also run treating self-rated health as a continuous measure and as an ordinal measure that maintained all of the categories and the results were essentially the same (available on request).

[9] Certified field interviewers measured respondents' resting, seated systolic and diastolic blood pressures (mmHg) and pulse rate (beats/minute). Following a five-minute seated rest, three serial measurements were performed at 30-second intervals. SBP reflects the average of measures 2 and 3. When either the second or third measure was missing, the other single measure was used. In cases where both measures 2 and 3 were missing, the first measure was used.

indicator for respondents who had completed a four-year degree. Of the 10,853 respondents in the analysis sample, 33 percent had attained a four-year degree.[10] Given the age range of respondents at Wave 4 (24 to 34 years old), this is likely a slight underestimate of the percentage who will ultimately attain a college degree.

Table 1: Descriptive Statistics for Health Measures.

	Black Female (N = 1,681)		Black Male (N = 1,295)	
	No Deg	Degree	No Deg	Degree
Self-rated health	0.42	0.62	0.49	0.64
Systolic blood pressure	124.07	120.33	130.01	130.14
Body mass index	33.06	31.00	28.94	29.85
Daily smoker	0.14	0.04	0.26	0.06
	White Female (N = 4,135)		White Male (N = 3,742)	
Self-rated health	0.51	0.78	0.54	0.76
Systolic blood pressure	120.21	118.78	130.19	129.19
Body mass index	29.66	26.41	29.29	27.65
Daily smoker	0.37	0.10	0.38	0.09

Note: Based on 10 complete data sets. Weighted data.

Table 1 provides weighted descriptive statistics for each of the measures of health and health behaviors for respondents with and without college degrees across the subgroups defined by sex and race/ethnicity. There are substantial differences in the proportions of people reporting very good or excellent self-rated health and smoking daily between those with college degrees and those without across the groups. Among black and white, females and males, respondents with degrees were more likely to report being in very good or excellent health than respondents without degrees. The proportion of daily smokers is higher among respondents without a college degree across all of the groups. The differences in BMI and systolic blood pressure between

[10] This is the unweighted percentage. The weighted percentage is 31 percent.

people with and without a college degree are more muted, but still present for black females and white males and females. Among black males, those with a college degree have slightly higher systolic blood pressure and BMI than those without a college degree, though the difference is minimal.

Additional Covariates. It is important to adjust for a number of covariates when estimating the effect of education on health and health behaviors.[11] The covariates can be generally divided into three groups: (1) measures of family socioeconomic status, (2) measures of ability, and (3) measures of childhood and adolescent health and health behaviors. All three groups of covariates have well-established associations with educational attainment and with health and health behaviors. In addition, all of the models include region (West, Midwest, South, Northeast) and the age of the respondents at Wave 4. Descriptive statistics for all of the covariates are available in Appendix One.

One of the benefits of using Add Health data is that it includes a number of measures of family socioeconomic status gathered at Wave 1 when respondents were in high school. This analysis draws on mother's and father's education (ten category measures ranging from "8th grade or less" to "professional training beyond a 4-year degree"), logged family income, family structure, and an interviewer assessment of respondent's living environment. Family structure is a five category variable that

[11] Many of these covariates are likely to be subject to measurement error. The primary goal in using the covariates is to construct weights based on propensity scores of attaining a college. As such, interest does not center on interpreting the structural effects of the covariates on either attaining a college degree or the given health outcomes, which would be affected by measurement error. Measurement error in the covariates may still have an effect in the models for the health outcomes, though the effect is not likely to be large given that the weights by design render the covariates largely insignificant predictors of the outcomes.

differentiates respondents living with two biological parents, two parents with at least one non-biological, a single mother, a single father, or some other arrangement (Harris 1999). The measure of the respondent's living environment is constructed as the average of the responses to the following two questions completed by Add Health interviewers: (1) "How well kept is the building in which the respondent lives?" and (2) "How well kept are most of the buildings on the street?" Responses to both questions ranged from "very poorly kept (needs major repairs)" (1) to "very well kept" (4).

Cognitive ability is often proposed as a source of spuriousness in analyses of the effects of education on health (Gottfredson 2004; Hirschi and Gottfredson 1994). Recent work suggests that it is unlikely that cognitive ability unrelated to education can account for the beneficial effects of education on health (Link et al. 2008). Nonetheless, this analysis draws on two measures of ability to address this possibility. At Wave 1 Add Health administered an abbreviated version of the Peabody Picture Vocabulary Test (PPVT) that serves as a measure of verbal ability or scholastic aptitude (Dunn and Dunn 1981). In addition, at Wave 1 respondents reported their grades in four subjects from the most recent semester (English, Math, History or Social Science, and Science). Self-reported GPA was constructed as the average of the grades from the available subjects.

The third group of covariates includes measures of health and health behaviors from childhood and adolescence. At Wave 1 parents reported the birth weight of respondents. Birth weight has been associated with educational attainment and a range of adult health outcomes (Almond and Currie 2011; Dahly, Adair and Bollen 2009). Add Health did not include a clinical measure of systolic blood pressure at Wave 1, but it did

include self-rated health, self-reported measures of height and weight that can be used to construct BMI, and the number of cigarettes smoked in the last month. In addition, the analysis adjusts for a measure of adolescent physical activity and measures of adolescent drinking. The measure of physical activity is taken as the maximum number of days per week respondents report engaging in any of a series of activities.[12] The measure of drinking stems from the questions "During the past 12 months, on how many days did you drink alcohol?" with responses ranging from "never" (1) to "every day or almost every day" (6).

Analysis Strategy

Following the approach outlined by Morgan and Todd (2008), the analysis to detect variation in the effects of education on health behaviors and outcomes proceeds in four steps for black and white, males and females. The first step in the analysis is to estimate a propensity score model. The second step is to use the propensity score models to construct weights that allow for estimates of different treatment effects. The third step is to estimate the average treatment effect (ATE), the average treatment effect of the treated (ATT), and the average treatment effect for the controls (ATC). The final step is to compare the estimates of the ATE, ATT, and ATC and make a determination whether they are all roughly the same or whether the ATT differs from the ATC. This section provides a brief description of each step.

[12] Wave 1 activities included: (a) roller-blading, roller-skating, skate-boarding, bicycling, (b) active sports such as baseball, softball, basketball, soccer, swimming, or football, (c) exercise such as jogging, walking, karate, jumping rope, gymnastics, or dancing.

Propensity score models are used to estimate the conditional probability of a "treatment" (Guo and Fraser 2010; Rosenbaum and Rubin 1983). For this analysis the "treatment" is attaining a college degree. The probability of attaining a college degree conditional on the covariates is estimated using a logistic regression model specified as

$$p_i = \Pr[D_i = 1|\mathbf{x_i}] = \frac{\exp(\mathbf{x_i}\phi)}{1+\exp(\mathbf{x_i}\phi)}, \quad (2.2)$$

where p_i is the propensity score for person i, D_i in an indicator for attaining a college degree, and $\mathbf{x}_i$ is a vector of covariates with coefficients $\boldsymbol{\phi}$. The covariates in the propensity score model should include all of the relevant predictors of the measures of health behaviors and outcomes (Brookhart et al. 2006; Guo and Fraser 2010; Rubin 1997). Square terms for mother's education, father's education, and logged family income were included in the propensity score model to capture their potential non-linear relationships with attaining a college degree and with health and health behaviors. The propensity score models were weighted by the sample weights provided by Add Health and estimated separately on each of the 10 imputed data sets.

The three groups of covariates, family SES, ability, and childhood and adolescent health and health behaviors, account for the main predictors of adult health that could be confounded with educational attainment. It is, however, likely that at least some relevant predictors have been omitted from the model. A sensitivity analysis is included to assess this concern.

The second step in the analysis involves calculating appropriate weights to obtain the various treatment effect estimators. The weight for the average treatment effect (ATE) is calculated as

$$w_{i,ATE} = \begin{cases} \dfrac{1}{\hat{p}_i} & \text{if } D_i = 1 \\ \dfrac{1}{1-\hat{p}_i} & \text{if } D_i = 0 \end{cases} . \tag{2.3}$$

Weighting by the inverse of the propensity score for the people who attained college degree and the inverse of one minus the propensity score for people who did not attain a college degree attempts to balance the sample with respect to the distribution of covariates between people who attained a college degree and people who did not. In other words, after weighting the data by the propensity scores, the distributions of the covariates among the people who attained a college degree should be similar to the distribution of the covariates among people who did not attain a college degree. In theory, this should hold for all moments of the distributions of the covariates, but in practice just the mean and the standard deviation are typically assessed.

One common metric to assess the balance of the means is the average of the standardized mean differences across the treatment and control groups (Morgan and Todd 2008; Rubin 1973). The standardized mean difference (StdMD) for a given variable is calculated as

$$\text{StdMD} = \frac{| \bar{x}_{i,D_i=1} - \bar{x}_{i,D_i=0} |}{\sqrt{\frac{1}{2} Var(x_{i,D_i=1}) + \frac{1}{2} Var(x_{i,D_i=0})}} . \tag{2.4}$$

Similarly, a common metric to assess the balance of the standard deviations is the average of the standardized standard deviation differences across the treatment and control groups. The standardized standard deviation difference (StdSD) for a given variable is calculated as

$$\text{StdSD} = \frac{|\, sd(x_{D_i=1}) - sd(x_{D_i=0}) \,|}{\sqrt{\frac{1}{2}Var(x_{i,D_i=1}) + \frac{1}{2}Var(x_{i,D_i=0})}}. \tag{2.5}$$

The StdSD is typically only used to assess balance among continuous covariates (Morgan and Todd 2008). In practice, perfect balance is never achieved and instead researchers assess how close the StdMD and StdSD are to 0 and how much balance is improved by the propensity score weight (Rubin 2006).

The weight for the average treatment effect for the treated (ATT) is calculated as

$$w_{i,ATT} = \begin{cases} 1 & \text{if } D_i = 1 \\ \dfrac{\hat{p}_i}{1-\hat{p}_i} & \text{if } D_i = 0 \end{cases}. \tag{2.6}$$

This weight treats the population-level group of people who attained a college degree as the target population by leaving the members of the sample who attained a college degree unweighted and weighting the members of the sample who did not attain a college degree according to their odds of having done so. This is an attempt to transform the sample of people who did not attain a degree into a representative sample of the population-level group of people who did attain a college degree (Morgan and Todd 2008). As with the ATE, the extent to which the weight balances the distribution of the covariates can be assessed.

Finally, the weight for the average treatment effect for the controls (ATC) is calculated as

$$w_{i,ATC} = \begin{cases} \dfrac{1-\hat{p}_i}{\hat{p}_i} & \text{if } D_i = 1 \\ 1 & \text{if } D_i = 0 \end{cases}. \tag{2.7}$$

Similar to the ATT, the purpose of this weight is to transform the sample into a representative sample of people who did not attain a college degree.

The propensity score weights have the same properties as other types of sample weights and can thus be combined in the standard way that different sample weights are combined (Kalton and Flores-Cervantes 2003). All of the weights outlined above were multiplied by the Add Health sample weights to incorporate the adjustments for unequal probability of sample selection and attrition into the estimates of the various treatment effects.

The third step is to estimate models for each of the measures of health and health behaviors to obtain the ATE, ATT, and ATC for attaining a college a degree for each of the subgroups defined by sex and race. Depending on the outcome, the models are either weighted linear regression models or weighted logistic regression models. The models take the general form

$$y_i = \alpha + \gamma D_i + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \tag{2.8}$$

where y_i is either a health measure or the logit of a health measure, D_i is an indicator for attaining a four-year college degree, γ gives the effect of a college degree with the precise interpretation depending on the weights, and $\mathbf{x}_i$ is a vector of covariates (the same covariates used in the propensity score model), and ε_i is a disturbance term. In theory the propensity score weights should balance the data such that it is unnecessary to adjust the covariates that went into the creation of the propensity scores when estimating the treatment effects. If, however, the propensity score model is misspecified, including the covariates in the regression models can help mitigate the effects of this misspecification (Bang and Robins 2005; Morgan and Winship 2007;

Robins and Rotnitzky 1997). The estimated standard errors are robust standard errors that adjust for heteroskedasticity and the clustering of respondents in schools in Wave 1. The standard errors do not specifically adjust for the additional uncertainty due to the estimation of the propensity scores; however, the robust standard errors should at least adjust for some of this uncertainty. Finally, the models are estimated on each of the imputed data sets and the results are combined according to Rubin's formulas using the multiple imputation suite of commands in Stata 12 (Little and Rubin 2002; StataCorp 2011).

The final step in the analysis is to assess whether the estimates of the ATT and ATC are different. Unfortunately, there is not a simple statistical test for the difference. It is, of course, possible to calculate the difference between the estimates, but it is not clear how to calculate an accurate standard error for the difference (Morgan and Todd 2008). Instead, this analysis relies on a substantive assessment of the magnitude of the differences and the consistency of the results across the different measures among black and white, females and males.

Results

The first step in the analysis is to estimate propensity scores for attaining a college degree for black and white, males and females. The parameter estimates for the propensity score models are provided in Appendix Two. Figure 1 illustrates the distribution of propensity scores for people who attained and did not attain a college degree for each subgroup based on the first complete data set.[13] Among black females

[13] The distributions were examined in all 10 of the complete data sets and were found to be quite similar.

and males, the propensity for attaining a college degree is relatively flat for those who did attain a degree and shows a sharp drop around 0.1 for those who did not attain a degree. White females and males who did not attain a degree have a similar propensity that peaks around 0.1 and then sharply drops afterward. In contrast, white females who did attain a degree have a steadily increasing propensity until around 0.9. There is a similar pattern for males, but the increase is not as pronounced.

Although the distributions of propensity scores are quite different for people who did and did not attain a degree, the distributions largely span the same range of propensities (see Table 2). The only particularly sparse region is among the highest decile of propensity scores. There are only a few people who did not attain a college degree, particularly among black males, with estimated propensity scores greater than or equal 0.9.[14]

[14] Even in the sparse regions, none of the complete data sets had 0 people in any cell.

Figure 1: Distribution of Estimated Propensity Scores.

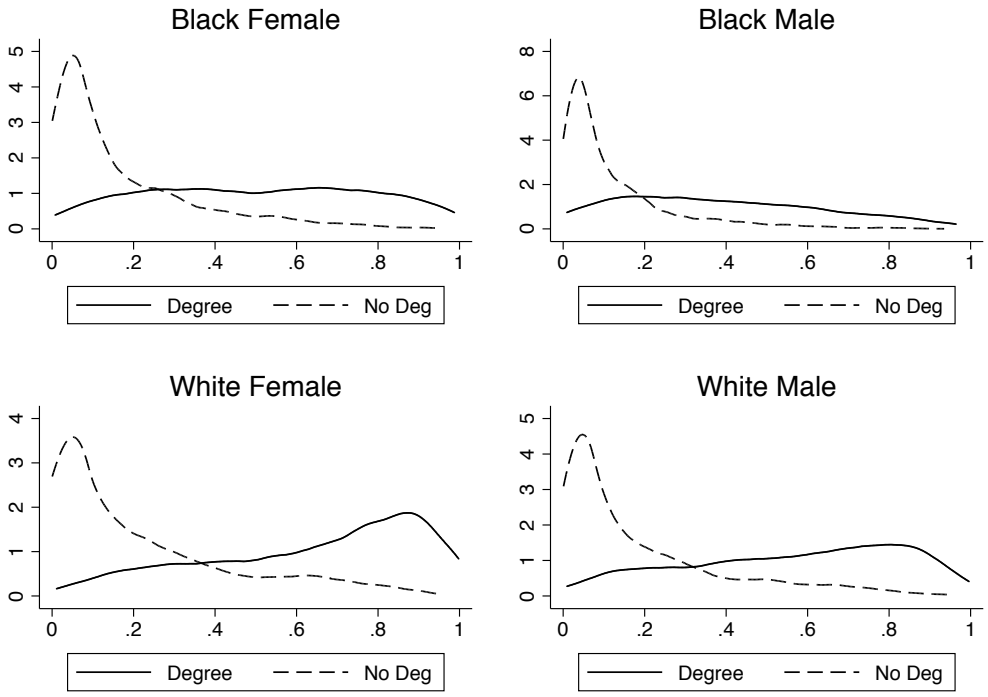


Table 2: Number of Cases in Deciles of Propensity Scores.

	Black Female		Black Male		White Female		White Male	
Prop.	No Deg	Deg	No Deg	Deg	No Deg	Deg	No Deg	Deg
[0.0 - 0.1)	500.4	32.2	582.7	40.8	1083.1	46.3	1237.9	59.5
[0.1 - 0.2)	225.8	42.8	189.8	35.2	467.8	77.2	471.6	84.5
[0.2 - 0.3)	134.0	54.5	94.0	36.7	300.4	114.2	295.8	98.8
[0.3 - 0.4)	86.4	50.3	56.0	27.2	200.7	122.8	184.5	110.9
[0.4 - 0.5)	74.4	56.9	35.1	28.1	142.7	135.2	124.8	114.8
[0.5 - 0.6)	52.7	56.1	23.5	26.3	114.3	144.4	107.5	126.9
[0.6 - 0.7)	32.4	66.8	14.4	22.5	104.6	175.8	81.4	146.5
[0.7 - 0.8)	23.6	55.5	10.1	25.7	70.5	217.5	52.6	155.8
[0.8 - 0.9)	10.0	66.2	8.7	21.5	47.7	289.0	27.2	154.5
[0.9 - 1.0]	2.3	57.7	1.7	15.0	10.2	270.6	7.7	98.8

Notes: Unweighted average counts across 10 complete data sets.

Given that all deciles have at least one case and most have many more cases, the main analysis proceeds using all of the cases representing the entire distribution of propensity scores. A sensitivity analysis in the following section assesses the robustness of the results with the sample limited to cases with greater overlap in the distribution of propensity scores.

The second step in the analysis is to construct the propensity score weights and assess the degree to which the weights balance the covariates. Table 3 provides estimates for the average standardized mean difference and average standardized standard deviation difference for the covariates using only the sample weights and the covariates using each of the propensity score weights. The distributions of the covariates differ substantially across people who did and did not attain college degrees for all of the groups as evidenced by the average difference in the standardized means. For instance, among black females, the average difference in the standardized means across all of the covariates is 0.356. Each of the propensity score weights substantially improves the balance in the covariates across the treatment and control groups with respect to the mean in all of the groups. For black females, the average difference in the standardized means across all of the covariates drops from 0.356 to 0.063 when applying the ATE weight, 0.039 when applying the ATT weight, and 0.091 when applying the ATC weight. The propensity score weights do less to improve balance with respect to the standard deviations of the covariates, particularly among black males where the ATE and ATC weights actually increase the average difference in the standardized standard deviation from 0.070 to 0.086. This suggests that the ATE and

ATC weights may not provide consistent estimates of the ATE and ATC for black males and caution is warranted in interpreting these results.

Table 3: Assessment of Mean and Standard Deviation Balance.

	Black Female		Black Male	
	Avg, StdMD	Avg, StdSD	Avg, StdMD	Avg, StdSD
Sample weight	0.356	0.084	0.337	0.070
ATE	0.063	0.060	0.084	0.086
ATT	0.039	0.038	0.037	0.035
ATC	0.091	0.068	0.101	0.103
	White Female		White Male	
	StdMD	StdSD	StdMD	StdSD
Sample weight	0.423	0.116	0.404	0.106
ATE	0.055	0.051	0.059	0.053
ATT	0.041	0.040	0.015	0.022
ATC	0.071	0.080	0.083	0.061

Notes: Balance metrics averaged over 10 complete data sets. ATE, ATT, and ATC incorporate sample weights.

Table 4 provides the estimates for the ATE, ATT and ATC along with a standard regression estimate (SRE) as a point of comparison for the two health outcomes and two measures of health behaviors.

Black Females. Attaining a college degree is significantly associated with an increased likelihood of reporting very good or excellent self-rated health, a substantial reduction in systolic blood pressure, and a decreased likelihood of being a daily smoker. The estimates from the standard regression models are quite similar to the estimates of the ATE using the propensity score weights. The largest difference between the standard regression estimate and the ATE is with effect of a college degree on smoking (SRE = -1.339, ATE = -1.875). The ATE is larger than the standard regression estimate, but it is still within two standard errors.

Among black females the ATC is greater (in absolute value) than the ATT for both self-rated health (ATT = 0.557, ATC = 0.751) and for smoking (ATT = -1.179, ATC = -2.223), while the ATT is greater than the ATC for BMI (ATT = -1.414, ATC = -0.813) and systolic blood pressure (ATT = -3.424, ATC = -3.114). The differences between the ATT and ATC for BMI and systolic blood pressure are not substantively large and the confidence intervals around the estimates allow for a high degree of overlap. The differences between the ATT and ATC for self-rated health and smoking are substantively large. In terms of the odds ratios (OR), for self-rated health the ATT OR is 1.75 compared to the ATC OR of 2.12, while for smoking the ATT OR is 0.31 compared to the ATC OR of 0.11. In addition, there is much less overlap in the confidence intervals of the estimates, particularly for smoking, than with systolic blood pressure. These results suggest that among black females there is evidence of variation in the health-related returns to a college degree and that returns would be greater for people who do not attain a degree were they to attain one than for those who attain a degree.

Black Males. Attaining a college degree is significantly associated with an increased likelihood of reporting very good or excellent health (SRE = 0.689) and a decreased likelihood of being a daily smoker (SRE = -1.510) based on the SRE. The ATE for smoking (ATE = -1.436) is similar to the SRE, but the ATE for self-rated health is not statistically significant. In addition, the ATT is greater (in absolute value) than the ATC for self-rated health and smoking. Given the issues with covariate balance among black males for the ATE and ATC as well as the large degree of overlap in the confidence intervals of the ATE and ATC for smoking, there is insufficient evidence to conclude that

the health-related returns to a college degree differ for black males who do and do not attain a college degree.

Table 4: Effects of a College Degree on Health.

	Black Female (N = 1,681)					
	SRE			ATE		
	Est	SE		Est	SE	
SRH	0.687	(0.187)	***	0.683	(0.222)	**
SBP	-3.291	(1.192)	**	-3.191	(1.549)	*
BMI	-0.835	(0.603)		-0.995	(0.724)	
Daily smoker	-1.339	(0.332)	***	-1.875	(0.377)	***
	ATT			ATC		
SRH	0.557	(0.205)	**	0.751	(0.257)	**
SBP	-3.424	(1.274)	**	-3.114	(1.796)	
BMI	-1.414	(0.644)	*	-0.813	(0.820)	
Daily smoker	-1.179	(0.343)	***	-2.223	(0.494)	***
	Black Male (N = 1,295)					
	SRE			ATE		
SRH	0.689	(0.251)	**	0.445	(0.302)	
SBP	0.242	(1.502)		0.492	(1.925)	
BMI	-0.101	(0.614)		-0.163	(0.623)	
Daily smoker	-1.510	(0.375)	***	-1.436	(0.374)	***
	ATT			ATC		
SRH	0.938	(0.259)	***	0.315	(0.336)	
SBP	0.420	(1.364)		0.404	(2.119)	
BMI	-0.491	(0.660)		0.007	(0.666)	
Daily smoker	-1.503	(0.401)	***	-1.484	(0.404)	***

Table 4 –continued-

	White Female (N = 4,135)					
	SRE			ATE		
	Est	SE		Est	SE	
SRH	0.664	(0.118)	***	0.600	(0.161)	***
SBP	-0.139	(0.637)		-0.123	(0.758)	
BMI	-1.283	(0.303)	***	-1.114	(0.333)	***
Daily smoker	-1.084	(0.365)	**	-1.092	(0.383)	**
	ATT			ATC		
SRH	0.711	(0.149)	***	0.547	(0.198)	**
SBP	-0.004	(0.863)		-0.126	(0.902)	
BMI	-1.084	(0.365)	**	-1.092	(0.383)	**
Daily smoker	-1.162	(0.151)	***	-1.075	(0.199)	***
	White Male (N = 3,742)					
	SRE			ATE		
SRH	0.487	(0.122)	***	0.381	(0.162)	*
SBP	-0.704	(0.634)		-0.256	(0.790)	
BMI	-0.896	(0.240)	***	-0.654	(0.272)	*
Daily smoker	-1.446	(0.166)	***	-1.252	(0.192)	***
	ATT			ATC		
SRH	0.493	(0.150)	***	0.325	(0.192)	
SBP	-0.025	(0.755)		-0.357	(0.989)	
BMI	-0.883	(0.293)	**	-0.548	(0.313)	
Daily smoker	-1.557	(0.180)	***	-1.153	(0.223)	***

Notes: Notes: Logit models used for self-rated health and daily smoker (reported estimates are log odds). Linear regression models used for systolic blood pressure and BMI. All results based on 10 complete data sets. Heteroskedastic and cluster robust standard errors are reported. All models include all covariates used in propensity score models.

White Females. Attaining a college degree is significantly associated with an increased likelihood of very good or excellent self-rated health (SRE = 0.664, ATE = 0.600), lower BMI (SRE = -1.283, ATE = -1.114), and a decreased likelihood of smoking daily (SRE = -1.084, ATE = -1.092). For both BMI and the likelihood of being a daily

smoker, the ATT and ATC estimates are roughly the same. For self-rated health, however, the ATT is greater than the ATC (ATT = 0.711, ATC = 0.547). The magnitude of the difference is substantively significant (ATT OR = 2.03 compared with ATC OR = 1.73), which suggests that for self-rated health, white females who attain a college degree benefit more from than degree than one would expect from white female who did not attain a degree were they to do so. This is consistent with the resource multiplication hypothesis.

White Males. There is a similar pattern of significant results among white males as was observed among white females. Attaining a college degree has a significant association with an increased likelihood of reporting very good or excellent self-rated health (SRE = 0.487, ATE = 0.381), reduced BMI (SRE = -0.896, ATE = -0.654), and a reduced likelihood of being a daily smoker (SRE = -1.446, ATE = -1.252). The estimates for the ATEs are generally smaller than SREs, but all of the estimates are within a standard error of each other. Among white males, the ATT estimates are consistently greater (in absolute value) than the ATC estimates (self-rated health: ATT = 0.493, ATC = 0.325; BMI: ATT = -0.883, ATC = -0.548; smoke: ATT = -1.557, ATC = -1.153). With the exception of smoking (ATT OR = 0.21, ATC OR = 0.32), the differences between the ATT and ATC are not particularly large; however, the pattern is consistent and the confidence intervals have relatively little overlap. This suggests that across several health outcomes, white males who attain a college degree accrue more health-related benefits than would white males who do not attain a college degree were they to do so.

Overall. Taken as a whole, the results suggest that among whites there is evidence that variation in the health-related returns to a college degree is greater for

people who attain a college degree than for people who do not attain a college degree were they to do so. This pattern of results is consistent with the resource multiplication hypothesis that whites are able to magnify other advantages related to health by attaining a college degree.

For blacks there is generally less evidence of variation in the health-related returns to a college degree. The one case where there appears to be some variation is in the effect of a college degree on smoking among black women where the ATC is greater than the ATT. This is consistent with resource substitution hypothesis, which suggests that attaining a college degree could be an important resource among black women to compensate for other health-related disadvantages. It is important to recall, however, that, particularly for blacks, the estimated propensity scores resulted in some sparse regions (i.e., there were few cases of people who did not attain a college degree but had a high propensity to do so) and the metrics assessing whether balance in the covariates was achieved were not ideal. The next section reports a series of sensitivity analyses to help determine the robustness of these results.

Sensitivity Analysis

The first sensitivity analysis involves only retaining people who have estimated propensity scores of attaining a college degree less than 0.8 for the analysis. Limiting the analysis sample to these people addresses the small number of cases with high estimated propensity scores who did not attain a college degree (particularly among blacks). Although this limits the generalizability of the results for people with high propensities for attaining a college degree, it provides a more robust basis for the estimated treatment effects for the remaining cases (roughly 90 percent of the analysis

sample). This analysis revealed the same pattern of results for black females, black males, and white males (results available upon request). For white females, the pattern of results is the same except for BMI. The ATT for BMI is estimated to be greater than the ATC (ATT = -1.177, ATC = -1.041), whereas in the main analysis the opposite pattern is observed. In both cases, however, the ATT and ATC are not substantively different and the same conclusion as in the main analysis is warranted. Limiting the sample to cases with propensity scores less than 0.8 does not alter the conclusions and suggests that the results do not hinge on the small number of cases with high propensity scores who did not attain a college degree.

As indicated in Table 3, there is room for improvement in the extent to which the estimated propensity score weights balance the data. Add Health contains a number of additional potential predictors of attaining a college degree that could be used in the propensity score models. Although in theory one should use the same predictors in the propensity score model as in the outcome model (Brookhart et al. 2006; Rubin 1997), it is possible that additional covariates if even weakly associated with the health outcomes could help reduce bias. The second sensitivity analysis reestimates the propensity scores using additional covariates.

The additional covariates used in the reestimation of the propensity scores included a number of measures of school behaviors from Wave 1 – number of times skipped school, number of grades repeated, ever received a suspension, and a scale consisting four measures related to creating problems in the classroom. The propensity score models also included college expectations, college aspirations, and attachment to school at Wave 1. These variables are all known to predict educational attainment,

though they are likely to have a weak association, at best, with health outcomes (especially conditional on the covariates already in the models).

Rerunning all of the analyses using the weights constructed from the updated propensity scores resulted in modest improvements in the balance metrics and the same substantive pattern of results for black females, white females, and white males (results available upon request). For black males, the balance metrics improved slightly, but in contrast to the main analysis, the ATC for daily smoking is estimated to be greater than the ATT for daily smoking in absolute value (ATT = -1.480, ATC = -2.224). The confidence intervals for the two estimates still show a significant degree of overlap and the ATC weights still do a relatively poor job of balancing the covariates. Therefore, the conclusion from the main analysis of no evidence of causal effect heterogeneity remains.

Overall, the two sensitivity analyses, restricting the sample to people with propensity scores less than 0.8 and adding additional covariates to the propensity score model, confirmed the results from the main analysis. It is, of course, still possible that the propensity score models suffer from misspecification or omitted variables, but the results of the sensitivity analyses suggest that these potential problems are not likely to be severe.

Discussion

This paper reports the results of an innovative approach to assessing variation in the health-related returns to a college degree. Past studies have found that the benefits of a college degree for health are unevenly distributed across racial and ethnic groups and sexes, over the life-course, and across cohorts. This analysis contributes to

our understanding of variation in the benefits of a college degree through an assessment of whether there is additional heterogeneity in the effects of higher education within a recent cohort of young adults among black and white, males and females. In addition to identifying variation in the health-related benefits of a college degree, the analytic strategy also allowed for an assessment of two competing theories of whether education alleviates or exacerbates health inequalities.

The results suggest that there is variation in the health-related returns to a college degree among white females and males. In particular, the benefits of higher education for health are greater for people who attained a degree than for people who did not attain a degree were they to do so. This is consistent with the resource multiplication theory and suggests that a college education contributes to widening health inequalities among advantaged and disadvantaged whites. Past research examining health inequalities has focused on disparities across rather than within racial and ethnic groups. These results indicate that it is also important to consider within group inequality.

The results suggest that there is also variation in the health-related returns to a college degree for black females. Higher education is associated with better self-rated health, lower systolic blood pressure, and a reduced likelihood of smoking among black women. Furthermore, there is evidence of variation in the returns to a college degree for self-rated health and smoking that is consistent with resource substitution theory. Among black women, the estimated health-related benefits of higher education are greater for people who did not attain a degree were they to do so than those who did attain a degree.

Finally, the results for black males are inconclusive. The diagnostics from the construction of the propensity score weights indicated that it is difficult to construct weights that balance the covariates for those who did and did not attain a degree. In particular, it is hard to identify black males who did not attain a college degree but had a high propensity to do so. This is most likely an indication that there are unobserved factors that differentiate black males who do and not attain college degrees. These results indicate not only that it is difficult to identify variation in the health-relate returns to higher education among black males, but that even the average estimate of the benefits of higher education is likely to be biased.

There are some limitations to this study worth keeping in mind that could be addressed in future research. First, the assessment of differences in the ATT and ATC and consequent determination of meaningful variation in the effects of a college degree relies on a substantive rather than statistical determination. Some of the differences in the ATT and ATC were deemed to be substantively meaningful, but it is unlikely they would be statistically significant were an appropriate test available. This suggests that more work is needed to see if these patterns replicate in other data sets and across other measures of health behaviors and outcomes.

Second, although the Add Health respondents show signs of future health problems, particularly related to obesity and hypertension, they are relatively young to experience significant health declines. Past research suggests that the health-related returns to education increase over the life course until a threshold is reached in old age, so it is possible that as the beneficial effects of education strengthen variation in the effects will also increase. If so, it will be easier to detect evidence of variation in the

health-related returns to education and, in particular, higher education. Future work should look to examine this possibility among older age groups.

Third, this study focused on a small number of health outcomes and health-related behaviors. The measures of health-related behaviors, BMI and smoking, are known to be among the leading behaviors causes of death in the US. One of the health outcomes, self-rated health, is known to be a reliable leading indicator of morbidity and mortality, while the other health outcome, systolic blood pressure, is closely related to cardiovascular disease. It would be useful, however, to extend the analysis to other health-related behaviors, particularly measures of diet and exercise, as well as other health outcomes, such as allostatic load. A consideration of these measures could help establish the robustness of the patterns observed in this analysis, particularly if assessed among older adults.

CHAPTER III: PROTECTIVE EFFECT OF A COLLEGE DEGREE FOR DEPRESSION

Depression is one of the leading causes of disability in the contemporary US, with close to 1 in 10 adults suffering from depressive symptoms (Strine et al. 2008; World Health Organization 2004). Numerous studies have documented a protective effect of education on depression (Bjelland et al. 2008; Kessler et al. 1994; Lorant et al. 2003; Mirowsky and Ross 2003; Turner and Lloyd 1999; Wheaton 1978). The protective effect of education, however, appears to vary substantially across different populations. The benefits of education in limiting depression are known to vary over the life-course (Bjelland et al. 2008; Miech and Shanahan 2000; Miech et al. 1999) and to be greater for women than for men (Ross and Mirowsky 2006). There is also evidence that there is no link between education and depression among blacks, particularly when depression is measured as the experience of major depressive episodes (Cockerham 1990; Hudson et al. 2012; Jackson and Cummings 2011; Williams et al. 2007).

This paper further contributes to our understanding of systematic variation in the protective effects of education on depression in three ways. First, the paper adopts a counterfactual framework for the analysis

(Heckman 2005; Morgan and Winship 2007; Rubin 1974; Splawa-Neyman [1923] 1990). This framework draws attention to the assumptions that underlie treating an estimated effect as a causal effect. To place the analysis in a counterfactual framework, a treatment needs to be defined. For this analysis, the treatment is attaining a college degree. Although past research measured educational attainment in years completed or as a series of transitions, the focus on attaining a college degree in this paper is strategic for two reasons. The attainment of a college degree currently represents the most significant educational threshold differentiating the life chances of people with and without degrees (Fischer and Hout 2006). In addition, the expansion in higher education has led to people attaining degrees who would not have in the past, which suggests that there may be more variation in general in the returns to a college degree among recent cohorts than one would expect among lower levels of education.

Second, the paper adopts an innovative analytic approach for detecting consequential variation, or heterogeneity, in treatment effects (Morgan and Todd 2008). The approach relies on assessing the difference between two treatment effects: the average treatment effect for the treated (ATT) and the average treatment effect for the controls (ATC). In this analysis, the ATT provides an estimate of how much a randomly chosen person who attained a college degree benefits from that degree with respect to depression. Similarly, the ATC provides an estimate of how much a randomly chosen person who did not attain a college degree would benefit from that degree with respect to depression if s/he were to attain a degree. A difference in the ATT and ATC provides evidence that the overall estimate of the protective effect of a college degree on depression, the average treatment effect (ATE), masks important dimensions of

heterogeneity (as will be detailed below) and provide a misleading or incomplete sense of the protective benefits of a college degree.

Third, the analysis draws on structural equation models (SEMs) to account for measurement error in the indicators of depression. There are two general approaches to the measurement of depression in analyses based on observational data: (1) the use of items based on depressive symptoms (sometimes referred to as dysthymia) and (2) the use of indicators of major depressive episodes (Murphy 2011). This analysis relies on the Center for Epidemiologic Studies Depression scale (CES-D), a commonly used instrument that measures depressive symptoms (Radloff 1977). Most analyses that use the CES-D treat it as a summative scale, but a recent examination of the psychometric properties of the CES-D indicates that it contains a mixture of different types of indicators and the indicators exhibit a substantial amount of measurement error (Perreira et al. 2005). Failing to account for measurement error in the indicators of depression can result in biased estimates of the protective effects of a college degree.

The data for this study come from Waves 1 and 4 of the National Longitudinal Study of Adolescent Health (Add Health), a nationally representative sample of young adults. At Wave 4, Add Health respondents are in their late 20s or early 30s, which is an ideal age range for this analysis because most respondents are old enough to have completed their education but have not yet settled into middle age, the period of the life course with the lowest prevalence of depression. All analyses are conducted separately for black and white, men and women.

Mechanisms Linking Education and Depression

Researchers have identified a number of mechanisms that could account for the protective effects of higher education on depression. First, educational attainment, and particularly attaining a higher degree, plays a central role in the stratification system in the contemporary U.S. (Blau and Duncan 1967; Fischer and Hout 2006; Hout and DiPrete 2006). People with a college degree are able to obtain more economic resources, which can be used to support their psychological wellbeing through adopting healthier lifestyles. In addition, the economic resources that typically accrue to people with a college degree can mitigate the chronic stress associated with poverty and economic hardship (Lantz et al. 2005; Turner, Wheaton and Lloyd 1995).

A higher degree, however, provides more than just economic resources. The human capital that develops with a college education can also be used to stave off depression by providing people with knowledge, habits, and attitudes that promote a sense of control over their lives (Mirowsky and Ross 1998; Mirowsky and Ross 2003; Ross and Mirowsky 1989). People who have a strong sense of control over their lives tend to have lower rates of depression, and conversely, a more fatalistic outlook is associated with higher levels of psychological distress (Turner and Lloyd 1999; Wheaton 1980).

Variation in the Protective Effect of a College Degree

Over the course of the 20th century college enrollment in the U.S. grew from a little over 200,000 people in 1900 to more than 15 million in 2000 (Snyder and Dillow 2011). In percentage terms, about a tenth of a percent of the population aged 15 to 24 were enrolled in college in 1900 as compared with roughly 35 percent of the population

aged 15 to 24 in 2000. This expansion in higher education has had profound implications for the life chances of people with and without college degrees (Fischer and Hout 2006). As higher education continues to expand, however, some of the people who currently attain degrees would most likely not have in the past. In addition, the recent growth in enrollment has been more concentrated among lower tiered colleges and universities than in the past. The changes in who is attaining a degree and the composition of institutions granting degrees has a potential impact on the protective effects of a college degree through both of the mechanisms outlined in the last section. If the returns to college degree on the labor market are not as consistent as in past cohorts, then attaining a college degree may not provide the economic resources to free individuals from the chronic stress associated with economic hardship. In addition, if more recent cohorts are not gaining as much human capital from their degree, then they may not also realize the same benefits with respect to a sense of control that protects against depression. These considerations raise the possibility that there may be more variation in the benefits of a college degree in recent cohorts.

Methodological Issues in Assessing Protective Effects

The standard approach to estimating the protective effect of education on depression is to regress a measure of depression, often a scale such as the CES-D, on a measure of education while adjusting for covariates thought to predict both depression and educational attainment. If the measure of education is attaining a college degree, then the model takes the general form

$$y_i = \alpha + \gamma D_i + \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \tag{3.1}$$

where y_i is a measure of depression, D_i is an indicator for attaining a college degree, $\mathbf{x}_i$ is a vector of covariates that are associated with both depression and the probability of attaining a college degree, and ε_i is a disturbance term. If one assumes that there are no omitted variables (i.e., $\text{Cov}(D_i, \varepsilon_i) = 0$ and $\text{Cov}(\mathbf{x}_i, \varepsilon_i) = 0$), then $\hat{\gamma}$ provides an estimate of the average causal effect of attaining a college degree on a given health outcome.

The standard approach has a straightforward interpretation in the counterfactual framework (Heckman 2005; Morgan and Winship 2007; Rubin 1974). Attaining a college degree can be thought of as a "treatment" – i.e., something that in principle can be manipulated (Holland 1986). The protective benefits for depression that arise from receiving the treatment, attaining a college degree, may vary from person to person. It is not, in general, possible or desirable to estimate a separate causal effect from a treatment for all individuals. Instead, analysts are typically interested in an average causal effect defined over a given population or subpopulation (Heckman 2005).

Identifying Variation in the Protective Effect

People who receive the treatment and people who do not receive the treatment are two common subpopulations of interest in the counterfactual framework. In this analysis the treatment is the attainment of a college degree, so the two subpopulations are people who did and did not attain a degree. The average causal effect for people who attained a college degree, the "average treatment effect for the treated" (ATT), is an estimate of how much a randomly selected person who attained a college degree benefited from that degree with respect to depression (Heckman 2005; Morgan and Harding 2006). Similarly, the average causal effect for people who did not attain a

college degree, the "average treatment effect for the controls" (ATC), is an estimate how much a randomly selected person who did not attain a college degree would benefit if she were to attain one (Heckman 2005; Morgan and Harding 2006). A comparison of the ATT and ATC provides a means of assessing the extent of heterogeneity in the protective effect of a college degree on depression.

There are two potential sources of heterogeneity in the protective effect of a college degree on depression that could lead to different estimates for the ATT and ATC (Morgan and Todd 2008; Morgan and Winship 2007). First, a difference in the ATT and ATC could arise if the benefits of a college degree depend on one of the other covariates in the model. In other words, there is an interaction effect between the treatment, attaining a college degree, and at least one of the other predictors of depression. Second, a difference in the ATT and ATC could arise if the effect of a college degree depends on an unobserved variable that predicts attaining a college degree. Methodologists divide causal effect heterogeneity due to unobserved variables into two types: (1) heterogeneity due to an omitted variable and (2) heterogeneity due to the possibility that individuals anticipate the potential gains from the treatment (or control) and select into the treatment (or control) to realize these gains (Heckman 2005; Heckman, Urzua and Vytlacil 2006; Morgan and Todd 2008; Morgan and Winship 2007). A comparison of the ATT and the ATC allows for a more complete assessment of potential variation in the protective effects of college degree and avoids the "curse of dimensionality" that comes with attempting to include all plausible interactions with observed variables.

Data

The analysis draws on Waves 1 and 4 from the National Longitudinal Study of Adolescent Health (Add Health).[15] The first wave of Add Health was based on a nationally representative sample of youth in grades 7 through 12 in the United States in 1994. The fourth wave of data collection occurred between January 2008 and February 2009, roughly 14 years after the first wave. About 75 percent of the 20,745 youth interviewed at Wave 1 were re-interviewed at Wave 4, resulting in 15,701 respondents with data from both waves.

Past work has found significant differences in the prevalence of depression by sex and race/ethnicity (Keyes and Goodman 2006; Riolo et al. 2005; Somervell et al. 1989). In addition, there are differences in the effect of socioeconomic status, and education in particular, on depression by sex and race/ethnicity (Hudson et al. 2012; Ross and Mirowsky 2006; Williams et al. 2007). Because of this, the analysis sample was restricted to non-Hispanic white and non-Hispanic black, males and females (N = 11,529) and separate models were estimated for the four subgroups. Add Health provides sample weights to adjust for the unequal probabilities of inclusion in the original sample frame and attrition between Waves 1 and 4. Respondents missing sample weights were excluded from the analysis sample (N = 674). Finally, two respondents missing information about the highest educational degree completed were dropped. This results in an analysis sample of 10,853 respondents.

---

[15] See Harris, Halpern, Whitsel, Hussey, Tabor, Entzel, and Udry Harris, Kathleen Mullan, Carolyn T. Halpern, Eric A. Whitsel, Jon M. Hussey, Joyce W. Tabor, Pamela P. Entzel, and J. Richard Udry. 2009. "The National Longitudinal Study of Adolescent Health: Research Design." URL: http://www.cpc.unc.edu/projects/addhealth/design. for a detailed description of the construction of the Add Health sample.

The remaining missing data were addressed using multiple imputation (Little and Rubin 2002). Ten complete data sets were constructed using the chained equation approach implemented in Stata 12 (StataCorp 2011). Most of the measures used in the analysis, including all of the depression indicators, were missing for less than 4 percent of the cases. The exceptions were some of the measures related to parental SES, particularly family income.[16] With the exception of a small number of outliers (i.e., less than 10), the range of the imputed values was quite similar to the range of the non-imputed values for all of these variables. In addition, the mean and variance of the measures for the non-imputed cases were quite similar (i.e., within a couple hundredths of a point) to the mean and variance of the measures with the imputed cases included.

Variables

Depression Items. Wave 4 respondents were asked a series of 10 questions taken from the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff 1977). Each item began with the stem "During the past seven days:" and included possible responses ranging from (0) "never or rarely" to (3) "most of the time or all of the time." Table 5 provides a list of the items, an abbreviation, and notes whether the indicator is best treated as an effect indicator, a causal indicator, or an outcome based on Perreira et al. (2005). Effect indicators change in response to changes in a latent variable, while changes causal indicators lead to changes in a latent variable. Both are valid indicators of latent variables, but they require different model specifications and have some

[16] Auxiliary variables that could improve the handling of missing data were considered, but none met the required associations outlined in chapter 4 to have an impact on reducing bias.

different properties (see Bollen and Bauldry (2011) for a more detailed discussion of the differences across types of indicators).

Table 5: CES-D Measures in Add Health.

Abbrev.	Item	Type
bother	You were bothered by things that usually don't bother you.	causal
blues	You could not shake off the blues, even with help from your family and friends.	effect
good	You felt you were just as good as other people.	causal
mind	You had trouble keeping your mind on what you were doing.	outcome
depress	You felt depressed.	effect
tired	You felt that you were too tired to do things.	causal
happy	You felt happy.	effect
enjoy	You enjoyed life.	causal
sad	You felt sad.	effect
dislike	You felt that people disliked you.	causal

Notes: The different types of indicators are discussed in the statistical analysis section.

College Education. In Wave 4 respondents were asked the highest educational level of education they had achieved to date. This information was used to construct an indicator for respondents who had completed a four-year degree. Of the 10,853 respondents in the analysis sample, 33 percent had attained a four-year degree.[17] Given the age range of respondents at Wave 4 (24 to 34 years old), this is likely a slight underestimate of the percentage who will ultimately attain a college degree.

Additional Covariates. The propensity score models and models estimating the effects of a college degree include a number of additional covariates that are related to

[17] This is the unweighted percentage. The weighted percentage is 31 percent.

both educational attainment and depression.[18] The covariates can be divided into three groups: (1) demographic measures, (2) measures of socioeconomic status in childhood, and (3) measures of ability (descriptive statistics for all the covariates are provided in Appendix Three). The demographic measures include the age of the respondents at Wave 4 and the region of the U.S. in which the respondent lived at Wave 1 (West, Midwest, South, Northeast).

Add Health contains a number of measures of family socioeconomic status gathered at Wave 1 when respondents were in high school. This analysis draws on mother's and father's education (ten category measures ranging from "8th grade or less" to "professional training beyond a 4-year degree"), logged family income, family structure, and an interviewer assessment of respondent's living environment. Family structure is a five category variable that differentiates respondents living with two biological parents, two parents with at least one non-biological, a single mother, a single father, or some other arrangement (Harris 1999). The measure of the respondent's living environment is constructed as the average of the responses to the following two questions completed by Add Health interviewers: (1) "How well kept is the building in which the respondent lives?" and (2) "How well kept are most of the buildings on the street?" Responses to both questions ranged from "very poorly kept (needs major repairs)" (1) to "very well kept" (4).

[18] Many of these covariates are also likely to contain measurement error. This is less of an issue than among the indicators of depression, however, because the covariates are primarily used in the construction of the propensity score weights where obtaining unbiased estimates of their structural effects is less of a concern than obtaining a good predicted probability of attaining a college degree. Measurement error in the covariates may still have an effect in the models for depression, though the effect is not likely to be large given that the weights by design render the covariates largely insignificant predictors of depression.

This analysis draws on two measures of ability. At Wave 1 Add Health administered an abbreviated version of the Peabody Picture Vocabulary Test (PPVT) that serves as a measure of verbal ability or scholastic aptitude (Dunn and Dunn 1981). In addition, at Wave 1 respondents reported their grades in four subjects from the most recent semester (English, Math, History or Social Science, and Science). Self-reported GPA was constructed as the average of the grades from the available subjects.

Analysis Strategy

The purpose of this analysis is to assess whether there is evidence of variation in the protective effects of a college degree for depression separately among black and white, females and males while accounting for measurement error in the indicators of depression. Several steps are needed in order to accomplish this: (1) measurement models need to be developed for depression, (2) propensity score models need to be estimated to form propensity score weights, (3) estimates need to be obtained for the ATT and ATC, and (4) a determination of whether the ATT and ATC differs needs to be made.

Measurement Models for Depression

The first step in the analysis is to develop models for depression that account for measurement error among white and black, men and women. The measurement models relate each of the potential indicators of depression to an underlying latent variable. This analysis considers two specifications of the measurement models informed by prior work on the CES-D items (Perreira et al. 2005). Perreira et al. (2005) distinguish three types of relationships the CES-D indicators may have with an

underlying latent variable for depression. The CES-D indicators may be effect indicators (indicators that change in response to changes in the underlying latent variable) or causal indicators (indicators that can lead to changes in the underlying latent variable). A few of the items in the CES-D scale appear to be best conceptualized as outcomes rather than indicators of depression.

Perreira et al. (2005) analyzed the psychometric properties of the CES-D items for an adolescent population and used a more complete set of items than is available in Add Health Wave 4, so it is not clear whether their results will hold for this analysis. Therefore, the development of the measurement models begins with a specification that treats all of the available indicators of depression as effect indicators. This is the implicit specification that underlies all of the analyses that create a summative scale and assess reliability with Cronbach's alpha (Bollen 1989). The second specification restricts the model to include only the measures found to be effect indicators by Perreira et al. (2005): depress, blues, happy, and sad. All of the measurement models are estimated separately for black and white, females and males using the Add Health sample weights on each of the 10 complete data sets and the results are combined using Rubin's rules (Asparouhov and Muthén 2010; Little and Rubin 2002; Meng and Rubin 1992).[19]

The two specifications are evaluated based on overall model fit statistics and the adequacy of the components of model fit. The overall model fit statistics include the chi-square test (a non-significant p-value indicates the model is consistent with the data),

[19] The measurement models will differ a bit when the propensity score weights are used rather than just the Add Health sample weights. Preliminary analyses examined variation in the measurement models using each of the different weights discussed in the next section and found no substantively meaningful differences.

the BIC (a value less than 0 favors the specified model over the saturated model; a decline in the BIC of 10 or more strongly favors the second over the first model), the CFI and TLI (values close to 1 indicate good model fit), and the RMSEA (values less than 0.05 indicate good model fit) (Bollen 1989; Browne and Cudeck 1993; Raftery 1995; Schwartz 1978; Steiger and Lind 1980). The components of model fit include the factor loadings and the amount of variance in the indicators explained by the latent variable (Bollen 1989). The chi-square test statistics (and standard errors) are adjusted for the use of weights using Mplus' robust maximum likelihood estimator, which provides a statistic that is asymptotically equivalent to the Yuan-Bentler chi-square test statistic (Muthén and Muthén 1998-2010; Yuan and Bentler 2000).

Estimating Propensity Score Weights

The second step of the analysis is to estimate propensity score models and construct weights for the various treatment effects. The estimation of propensity scores accomplishes two things in this analysis. First, the propensity scores are used to construct weights that can be used to estimate the ATT and the ATC. Second, a propensity score approach in general draws attention to the extent to which adjusting for the covariates creates comparable groups of people with and without the treatment (a college degree). As will be discussed below, this ends up being a significant issue in this analysis, which would not be apparent in a standard regression analysis.

Propensity score models are used to estimate the conditional probability of a "treatment" (Guo and Fraser 2010; Rosenbaum and Rubin 1983). For this analysis the "treatment" is attaining a college degree. The probability of attaining a college degree conditional on the covariates is estimated using a logistic regression model specified as

$$p_i = \Pr[D_i = 1|\mathbf{x_i}] = \frac{\exp(\mathbf{x_i}\phi)}{1+\exp(\mathbf{x_i}\phi)}, \quad (3.2)$$

where p_i is the propensity score for person i, D_i in an indicator for attaining a college degree, and $\mathbf{x}_i$ is a vector of covariates with coefficients $\boldsymbol{\phi}$. The covariates in the propensity score model should include all of the relevant predictors of health and health behaviors (Brookhart et al. 2006; Guo and Fraser 2010; Rubin 1997). Square terms for mother's education, father's education, and logged family income were included in the propensity score model to capture potential non-linear relationships with attaining a college degree and with depression. The propensity score models were weighted by the sample weights provided by Add Health and estimated separately for black and white, males and females on each of the 10 complete data sets.

The propensity score models are used to calculate appropriate weights to obtain the various treatment effect estimators. The weight for the average treatment effect (ATE) is calculated as

$$w_{i,ATE} = \begin{cases} \dfrac{1}{\hat{p}_i} & \text{if } D_i = 1 \\ \dfrac{1}{1-\hat{p}_i} & \text{if } D_i = 0 \end{cases}. \quad (3.3)$$

Weighting by the inverse of the propensity score for the people who attained a college degree and the inverse of one minus the propensity score for people who did not attain a college degree attempts to balance the sample with respect to the distribution of covariates between people who attained a college degree and people who did not. In other words, after weighting the data by the propensity scores, the distributions of the covariates among the people who attained a college degree should be similar to the

distribution of the covariates among people who did not attain a college degree. In theory, this should hold for all moments of the distributions of the covariates, but in practice just the mean and the standard deviation are typically assessed.

One common metric to assess the balance of the means is the average of the standardized mean differences across the treatment and control groups (Morgan and Todd 2008; Rubin 1973). The standardized mean difference (StdMD) for a given variable is calculated as

$$\text{StdMD} = \frac{|\,\overline{x}_{i,D_i=1} - \overline{x}_{i,D_i=0}\,|}{\sqrt{\frac{1}{2}Var(x_{i,D_i=1}) + \frac{1}{2}Var(x_{i,D_i=0})}}. \tag{3.4}$$

Similarly, a common metric to assess the balance of the standard deviations is the average of the standardized standard deviation differences across the treatment and control groups. The standardized standard deviation difference (StdSD) for a given variable is calculated as

$$\text{StdSD} = \frac{|\,sd(x_{D_i=1}) - sd(x_{D_i=0})\,|}{\sqrt{\frac{1}{2}Var(x_{i,D_i=1}) + \frac{1}{2}Var(x_{i,D_i=0})}}. \tag{3.5}$$

The StdSD is only used to assess balance among continuous covariates (Morgan and Todd 2008). For both the StdMD and the StdSD perfect balance is achieved when the metrics equal 0, which indicates that there is no difference between people who attained a college degree and people who did not in the means or the standard deviations for any of the covariates when the data is weighted by the ATE. In practice, perfect balance is never achieved and instead researchers assess how close the StdMD and StdSD are to 0 and how much balance is improved by the propensity score weight.

The weight for the average treatment effect for the treated (ATT) is calculated as

$$w_{i,ATT} = \begin{cases} 1 & \text{if } D_i = 1 \\ \dfrac{\hat{p}_i}{1-\hat{p}_i} & \text{if } D_i = 0 \end{cases} . \tag{3.6}$$

This weight treats the population-level group of people who attained a college degree as the target population by leaving the members of the sample who attained a college degree unweighted and weighting the members of the sample who did not attain a college degree according to their odds of having done so. This is an attempt to transform the sample of people who did not attain a degree into a representative sample of the population-level group of people who did attain a college degree (Morgan and Todd 2008). As with the ATE, the extent to which the weight balances the distribution of the covariates can be assessed.

Finally, the weight for the average treatment effect for the controls (ATC) is calculated as

$$w_{i,ATC} = \begin{cases} \dfrac{1-\hat{p}_i}{\hat{p}_i} & \text{if } D_i = 1 \\ 1 & \text{if } D_i = 0 \end{cases} . \tag{3.7}$$

Similar to the ATT, the purpose of this weight is to transform the sample into a representative sample of people who did not attain a college degree.

The propensity score weights have the same properties as other types of sample weights and can thus be combined in the standard way that different sample weights are combined (Kalton and Flores-Cervantes 2003). All of the weights outlined above were multiplied by the Add Health sample weights to incorporate the adjustments for

unequal probability of sample selection and attrition into the estimates of the various treatment effects.

Estimating Treatment Effects

The third step is to estimate weighted SEMs that incorporate the measurement model for depression and the structural model to obtain the ATE, ATT, and ATC for attaining a college degree in each of the subgroups defined by sex and race. The SEMs can be represented by the following system of equations

$$\eta_i = \alpha_\eta + \gamma D_i + \boldsymbol{\beta}'\mathbf{x}_i + \zeta_i \tag{3.8}$$

$$y_{ij} = \alpha_j + \lambda_j \eta_i + \varepsilon_{ij}, \tag{3.9}$$

where y_{ij} is j^{th} indicator of depression, η_i is a latent variable for depression, D_i is an indicator for attaining a four-year college degree, γ gives the effect of a college degree with the precise interpretation depending on the weights, $\mathbf{x}_i$ is a vector of covariates (the same covariates used in the propensity score model), and ζ_i and ε_{ij} are disturbance terms. The models assume that the disturbance terms are all uncorrelated. The scale for latent depression is set by fixing the factor loading to 1 and the intercept to 0 for depress (the a priori most reliable indicator).

In theory the propensity score weights should balance the data such that it is unnecessary to adjust the covariates that went into the creation of the propensity scores when estimating the treatment effects. If, however, the propensity score model is misspecified, including the covariates in the regression models can help mitigate the effects of this misspecification (Bang and Robins 2005; Morgan and Winship 2007; Robins and Rotnitzky 1997). The estimated standard errors are robust standard errors that adjust for heteroskedasticity and the clustering of respondents in schools in Wave

1. The standard errors do not adjust for the additional uncertainty due to the estimation of the propensity scores; however, the robust standard errors should at least adjust for some of this uncertainty. Finally, the models are estimated on each of the 10 complete data sets and the results are combined according to Rubin's formulas using the multiple imputation suite of commands in Stata 12 (Asparouhov and Muthén 2010; Little and Rubin 2002; Meng and Rubin 1992; StataCorp 2011).

Assessing Consequential Effect Heterogeneity

The final step in the analysis is to assess whether the estimates of the ATT and ATC are different. Unfortunately, there is not a simple statistical test for the difference. It is, of course, possible to calculate the difference between the estimates, but it is not clear how to calculate an accurate standard error for the difference (Morgan and Todd 2008). Instead, this analysis relies on a substantive assessment of the magnitude of the differences within the sex and race subgroups.

Results

Table 6 presents the overall model fit statistics for the two different specifications of the measurement models for black and white, males and females. The purpose of the measurement model is to relate the indicators of depression to a latent variable for depression that can be used as the outcome in the following analysis. The measurement model that relies on just the four effect indicators (depress, blues, happy, and sad) provides a better fit with the data than the measurement model that includes all of the indicators for all of the groups. Among black females and males, the χ^2 test statistic is not significant, the BIC is negative, the CFI and TLI are close to 1, and the

RMSEA is below 0.05 for the model with four effect indicators, all of which indicate a good fit with the data. Among white males all of the measures of model fit except for the χ^2 test statistic indicate a good fit with the data. The measurement model with four effect indicators for white females does not fit the data as well as the same model for the other groups. Although the CFI and TLI are close to 1 and the RMSEA is right at the 0.05 threshold, the χ^2 test statistic is significant and the BIC is greater than 0. There are, however, no obvious theoretically informed alternative specifications for the model, so the four-indicator measurement model is maintained for the following analyses.

Table 6: Overall Model Fit Statistics for Depression Models.

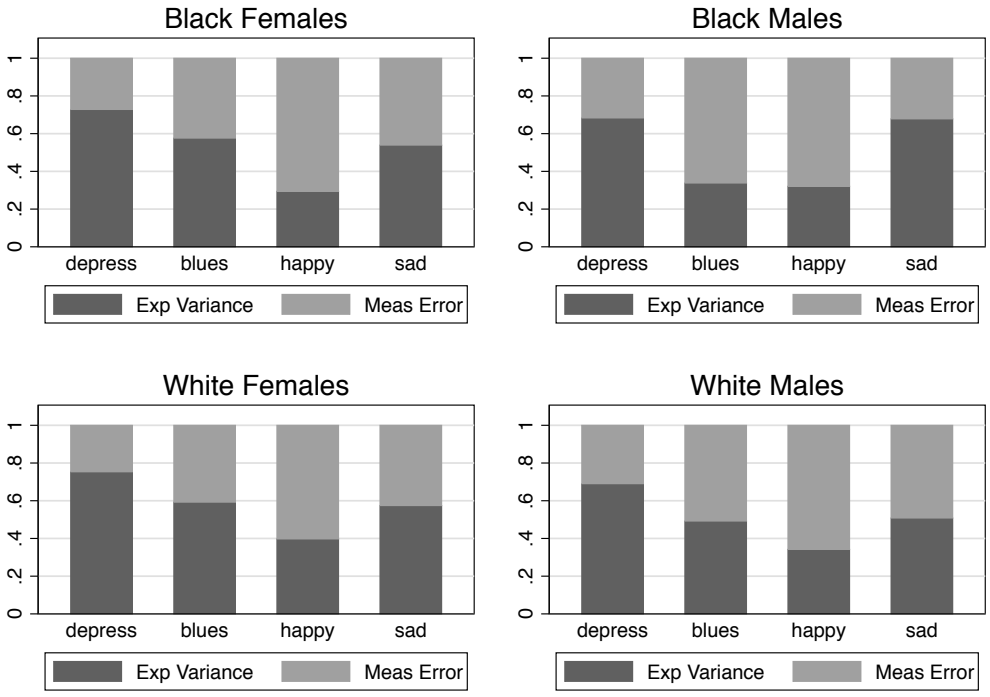
	ChiSq	df	p-value	BIC	CFI	TLI	RMSEA
Black Females (N = 1681)							
All indicators	339.005	35	0.000	79.055	0.829	0.780	0.072
4 effect indicators	4.045	2	0.132	-10.809	0.996	0.989	0.025
Black Males (N = 1295)							
All indicators	268.284	35	0.000	17.465	0.826	0.777	0.072
4 effect indicators	1.511	2	0.470	-12.822	1.000	1.004	0.000
White Females (N = 4135)							
All indicators	1226.834	35	0.000	935.381	0.792	0.733	0.091
4 effect indicators	25.768	2	0.000	9.114	0.986	0.958	0.054
White Males (N = 3742)							
All indicators	887.895	35	0.000	599.937	0.817	0.764	0.081
4 effect indicators	8.591	2	0.014	-7.864	0.995	0.986	0.030

Notes: Based on 10 complete data sets. Weighted using Add Health sample weight. The four effect indicators are depress, blues, sad, and happy.

The components of model fit also indicate the four-indicator measurement models are a good fit with the data. The factor loadings for the indicators are statistically significant in the expected direction across all of the sex and race subgroups. Figure 1 illustrates the proportion of the variance in each of the indicators

accounted for by latent depression and the remaining unaccounted for variance (typically considered measurement error). Latent depression accounts for a sizable proportion of the variance in all of the indicators across all the race and sex subgroups. Not surprisingly, the most reliable indicator (as determined by the highest explained variance) is depress and the least reliable is happy, the indicator with the opposite valence of the other three. Figure 2 also illustrates the non-negligible extent of measurement error in each of the items, which underscores the importance of accounting for this error in the analysis of the protective effects of a college degree.

Figure 2: Explained Variance and Measurement Error.



The next step in the analysis is to estimate propensity scores for attaining a college degree for black and white, males and females. The parameter estimates for the

propensity score models are provided in Appendix Four. It is useful to examine the distributions of the estimated propensity scores to check for regions of common support (i.e., ranges of probabilities where there are cases that both attained and did not attain a college degree). Figure 3 illustrates the distribution of the estimated propensity scores for people who attained and did not attain a college degree based on the first complete data set.[20] Among black females and males, the propensity for attaining a college degree is relatively flat for those who did attain a degree and shows a sharp drop around 0.1 for those who did not attain a degree. White females and males who did not attain a degree have a similar propensity that peaks around 0.1 and then sharply drops afterward. In contrast, white females who did attain a degree have a steadily increasing propensity until around 0.9. There is a similar pattern for males, but the increase is not as pronounced.

Although the distributions of propensity scores are quite different for people who did and did not attain a degree, the distributions largely span the same range of propensities (see Table 7). The only particularly sparse region is among the highest decile of propensity scores. There are only a few people who did not attain a college degree, particularly among black males, with estimated propensity scores greater than or equal 0.9.[21] Given that all deciles have at least one case and most have many more cases, the main analysis proceeds using all of the cases representing the entire distribution of propensity scores.

---

[20] The distributions of estimated propensity scores were examined in all 10 of the complete data sets and were found to be quite similar.

[21] Even in the sparse regions, none of the complete data sets had 0 people in any cell.

Figure 3: Distribution of Estimated Propensity Scores.

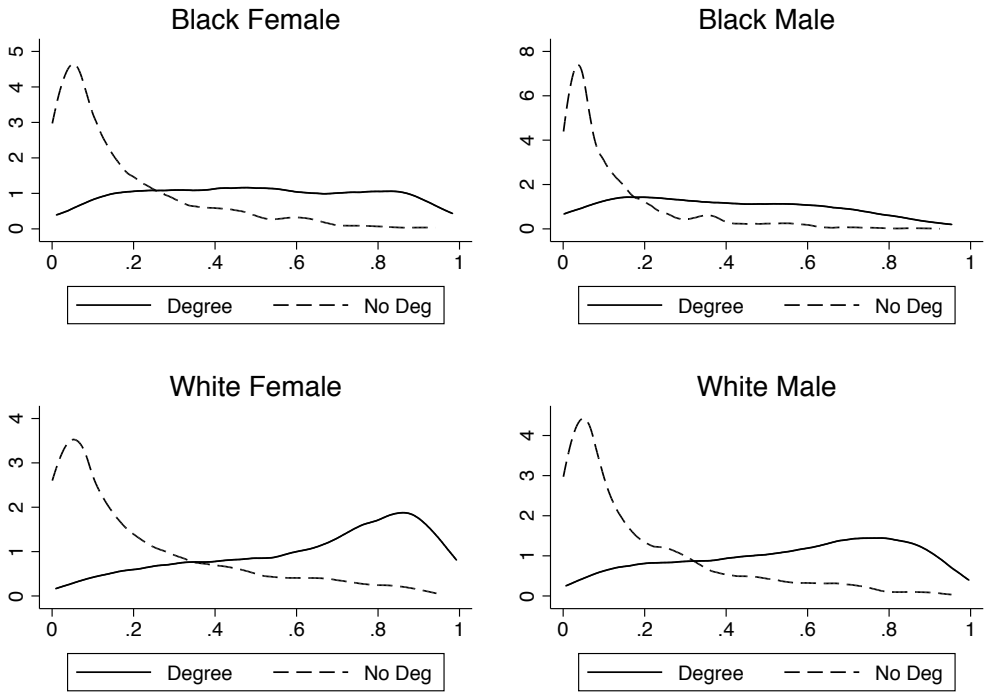


Table 7: Number of Cases in Deciles of Propensity Scores.

	Black Female		Black Male		White Female		White Male	
Propensity	No Dg	Degree	No Dg	Degree	No Dg	Degree	No Dg	Degree
[0.0 - 0.1)	473.3	29.9	583.1	40.6	1076.6	48.0	1217.5	59.8
[0.1 - 0.2)	233.9	42.0	193.8	38.0	461.5	71.1	482.5	85.8
[0.2 - 0.3)	140.0	55.5	94.6	35.4	306.9	114.6	300.6	95.4
[0.3 - 0.4)	101.9	57.4	56.7	30.8	196.6	130.2	188.1	113.0
[0.4 - 0.5)	72.2	55.8	30.1	27.8	147.4	138.8	129.8	123.4
[0.5 - 0.6)	51.8	58.7	24.7	22.4	114.6	146.9	101.5	122.0
[0.6 - 0.7)	34.3	61.3	15.8	27.9	100.6	177.1	88.9	144.2
[0.7 - 0.8)	21.5	56.1	8.8	24.2	74.9	212.5	47.9	153.4
[0.8 - 0.9)	9.8	72.5	7.2	22.1	52.7	292.3	26.1	162.4
[0.9 - 1.0]	3.3	49.8	1.2	9.8	10.2	261.5	8.1	91.6

Notes: Unweighted average counts across 10 complete data sets.

The next step in the analysis is to construct the propensity score weights and assess the degree to which the weights balance the covariates. Table 8 provides

estimates for the average standardized mean difference and average standardized standard deviation difference for the covariates using only the sample weights and the covariates using each of the propensity score weights. The propensity score weights are generally successful at improving the balance of the means of each of the covariates across the sex and race subgroups. For instance, the average standardized difference in the means for black females is 0.41 using just the Add Health sample weights. This average standardized difference is reduced to 0.072 when the ATE weight is applied, to 0.045 when the ATT weight is applied, and to 0.109 when the ATC weight is applied. This same pattern is observed across black and white, females and males. The propensity score weights are least successful balancing the covariates for black males.

In general, the propensity score weights also improve the balance of the covariates with respect to the average difference in the standard deviations. Again, considering black females, the average difference in the standard deviations across all of the covariates is 0.073. This difference drops to 0.051 when the ATE is applied, 0.034 when the ATT is applied, and 0.054 when the ATC is applied. The propensity score weights do not improve the balance with respect to the variance as much as they do with respect to the mean, but this is partially because the covariates are already reasonably balanced with respect to the variance across people who did and did not attain a college degree.

Although the propensity score weights are generally successful at improving the balance in the distributions of the covariates across people who did and did not attain college degrees, there is still room for improvement. Ideally, the StdMD and StdSD would be closer to 0. The fact that they are not suggests that either there is some

misspecification in the propensity score model or that it is not possible to achieve perfect balance in the observed data. This underscores the importance of using the doubly-robust estimator that adjusts for the covariates in the models that use the propensity score weights.

Table 8: Assessment of Mean and Standard Deviation Balance.

	Black Female		Black Male	
	StdMD	StdSD	StdMD	StdSD
Sample weight	0.410	0.073	0.377	0.064
ATE	0.072	0.051	0.128	0.094
ATT	0.045	0.034	0.031	0.041
ATC	0.109	0.054	0.149	0.110
	White Female		White Male	
	StdMD	StdSD	StdMD	StdSD
Sample weight	0.480	0.130	0.464	0.115
ATE	0.063	0.060	0.049	0.060
ATT	0.035	0.050	0.013	0.025
ATC	0.078	0.082	0.070	0.068

Notes: Balance metrics averaged over 10 complete data sets. ATE, ATT, and ATC incorporate sample weights.

The final step in the analysis is to estimate the various treatment effects and to make a determination as to whether there is heterogeneity in the protective effects of a college degree on depression. In addition to the models based on the three propensity score weights, results are reported for a model that just uses the Add Health sample weight. The model that uses just the Add Health sample weight reflects a standard regression estimate (SRE) of the effect of a college degree on depression while adjusting for covariates and accounting for measurement error. This model is intended as a point of comparison for the estimates of the treatment effects.

Table 9: Overall Model Fit Statistics for SEMs Estimating Treatment Effects.

	ChiSq	df	p-value	BIC	CFI	TLI	RMSEA
Black Female (N = 1681)							
SRE	80.597	59	0.032	-357.605	0.982	0.975	0.015
ATE	88.110	59	0.008	-350.092	0.963	0.948	0.017
ATT	90.613	59	0.005	-347.589	0.964	0.950	0.018
ATC	84.868	59	0.015	-353.334	0.960	0.944	0.016
White Female (N = 4135)							
SRE	115.487	59	0.000	-375.820	0.985	0.980	0.015
ATE	65.868	59	0.251	-425.439	0.997	0.995	0.005
ATT	95.052	59	0.002	-396.255	0.983	0.976	0.012
ATC	68.969	59	0.176	-422.338	0.993	0.991	0.006
Black Male (N = 1295)							
SRE	73.086	59	0.103	-349.724	0.981	0.973	0.013
ATE	106.322	59	0.000	-316.488	0.911	0.876	0.025
ATT	95.557	59	0.002	-327.253	0.908	0.872	0.020
ATC	120.581	59	0.000	-302.229	0.886	0.841	0.028
White Male (N = 3742)							
SRE	60.907	59	0.407	-424.508	0.999	0.999	0.003
ATE	224.201	59	0.000	-261.214	0.901	0.862	0.027
ATT	108.031	59	0.000	-377.384	0.964	0.949	0.015
ATC	293.813	59	0.000	-191.602	0.859	0.804	0.032

Notes: Based on 10 complete data sets.

Table 9 reports the overall model fit statistics for the four models based on different weights (SRE, ATE, ATT, and ATE weights) for black and white, female and males. The chi-square test remains significant for most of the models, but, with a couple exceptions, the other measures of overall model fit suggest the models have a decent fit with the data. The BICs are all well below 0, the CFIs and TLIs are generally greater than 0.95, and the RMSEAs are all below 0.05. The main exception is among black males. The

CFI and TLI for models using the three propensity score weights are low, especially for the model using the ATC weight (CFI = 0.886, TLI = 0.841). This is perhaps not surprising given the challenge of developing propensity score weights for black males (i.e., the sparse regions noted in Table 7 and difficulty achieving balance across the covariates noted in Table 8). Finally, the models using the ATE and ATC weights among white males also show signs of marginal fit with relatively low CFIs and TLIs.

Table 10: Estimates of the Effect of a College Degree on Depression.

	Est	SE	
Black Female (N = 1681)			
SRE	-0.156	(0.055)	**
ATE	-0.224	(0.056)	***
ATT	-0.124	(0.055)	***
ATC	-0.258	(0.064)	***
White Female (N = 4135)			
SRE	-0.105	(0.031)	***
ATE	-0.107	(0.047)	*
ATT	-0.119	(0.043)	**
ATC	-0.103	(0.064)	
Black Male (N = 1295)			
SRE	-0.137	(0.077)	
ATE	-0.224	(0.079)	**
ATT	-0.096	(0.066)	
ATC	-0.238	(0.087)	**
White Male (N = 3742)			
SRE	-0.058	(0.028)	*
ATE	-0.063	(0.040)	
ATT	-0.010	(0.025)	
ATC	-0.092	(0.052)	

Notes: Based on 10 complete data sets.

Table 10 reports the SRE, ATE, ATT, and ATC estimates of the protective effects of a college degree on depression (net of demographic characteristics, childhood SES,

and ability). Among black females all of the estimates show a statistically significant protective effect of a college degree (SRE = -0.156, ATE = -0.224, ATT = -0.124, ATC = -0.258). The estimate for the SRE is a bit lower than the ATE, but within a standard confidence interval. The ATC is around twice the magnitude of the ATT with little overlap in the confidence intervals, which suggests that black females who did not attain a college degree would realize a greater protective benefit from a degree than black females who did attain a degree.

For white females, the SRE, ATE, ATT, and ATC estimates of the protective effect of a college degree on depression are all essentially the same (SRE = -0.105, ATE = -0.107, ATT = -0.119, ATC = -0.103). The ATC is not statistically significant, but this appears to reflect the greater uncertainty in this estimate associated with the higher standard error. The similarity of all of the estimates suggests that there is little variation in the protective effect of a college degree for white women.

Among males, there is less evidence that a college degree protects against depression. The SRE estimate for white males is statistically significant and roughly the same magnitude as the ATE (SRE = -0.058, ATE = -0.063), but the size of the effect is about half that of white women (ATE = -0.107) and a quarter that of black women (ATE = -0.224). Black males have a statistically significant ATE and ATC that are close to the estimates of black females. These estimates, however, should be interpreted with caution given the challenges with the propensity scores and the marginally acceptable overall model fit.

Discussion

Depression is one of the leading causes of disability in the US. Numerous studies have documented the protective effects of education against depression, but the benefits of education vary systematically across groups. Past research has found that the relationship between education and depression varies across the life course, appears to be stronger for women than men, and may not exist at all among blacks. This study contributes to our understanding of whether there is additional heterogeneity in the beneficial effects of education within black and white, female and male subgroups. In addition, the analytic approach involving the use of propensity scores reveals an important issue in the assessment of the protective effects of education among black males.

Past research has found little evidence of a relationship between education and depression among blacks. This analysis, however, shows that a college degree appears to matter for black females, which suggests that analyses that examine the protective effects of education on depression for blacks overall are missing an important interaction with sex. In addition, adopting a counterfactual framework draws attention to the challenge in estimating the beneficial effect of a college degree on depression for black males. The non-significant estimate from the standard regression approach (while adjusting for measurement error) is consistent with recent findings among blacks in general. The challenges encountered with the construction of the propensity scores, however, suggest that it is not clear that a comparison of depression among black men with and without college degrees provides a fair estimate of the protective effect of a college degree. It is difficult to achieve balance with respect to other

covariates that may influence both educational attainment and depression among black males. This means that research relying on observational data needs to develop a richer set of covariates to account for the differences in black males who do and do not attain a college degree in order to get a more accurate estimate of the protective effects of higher education on depression for this group.

Past work has found that women realize a greater benefit from education than men with respect to depression (Ross and Mirowsky 2006). This pattern is found among whites in this analysis as white women are estimated to have roughly twice the protective effect of a college degree as white men. The pattern among blacks, however, is inconclusive given the challenges in estimating the protective effects of a college degree for black men. An interesting dynamic that has not been previously noted is that black women are estimated to gain between 50 and 100 percent more from a college degree than white women with respect to depression.

Finally, the analysis adopted an innovative approach to determine whether there is evidence of additional variation in the protective effects of a college degree on depression. Among black females, the estimates of the ATT and ATC suggest that there is consequential heterogeneity in the benefits of a college degree. In particular, black women who did not attain a college degree are estimated to gain more from a degree were they to complete one than the benefits black women who completed a degree receive. Contrary to expectations, there is no indication of additional variation in the benefits of a college degree among whites. For white women the estimates of the ATT and ATC are quite similar to the estimate of the ATE. For white men, there is a difference in the ATT and ATC, but neither are statistically significant.

There are some limitations to this study worth keeping in mind that could be addressed in future research. First, the assessment of differences in the ATT and ATC and consequent determination of meaningful variation in the effects of a college degree relies on a substantive rather than statistical determination. Some of the differences in the ATT and ATC were deemed to be substantively meaningful, but it is unlikely they would be statistically significant were an appropriate test available. This suggests that more work is needed to see if these patterns replicate in other data sets and across other measures of health behaviors and outcomes.

Second, this study relied on the CES-D measures of depressive symptoms, an often-used approach to measuring depression. Another approach is to analyze the occurrence of major depressive episodes, which are less commonly observed than depressive symptoms. Educational attainment has been linked to depression in analyses using both approaches, but future work should analyze whether there is evidence of similar patterns of variation in the protective effect of higher education when analyzing major depressive episodes.

Finally, this analysis focused on depression among a recent cohort of young adults. This was a strategic decision because there are good reasons to expect more variation in the returns to a college degree among recent cohorts. The prevalence of depression, and particularly major depressive episodes, however, are known to be higher among older adults as is the association between education and depression. As recent cohorts age and new data becomes available, future research should continue to examine the extent of variation in the protective effects of a college degree among older populations.

CHAPTER IV: AUXILIARY VARIABLES FOR ADDRESSING MISSING DATA

Social scientists increasingly rely on multiple imputation (MI) or direct maximum likelihood estimators (DML)[22] to address missing data in their analyses. Both of these approaches provide estimators with desirable properties under less restrictive assumptions than other approaches (e.g., complete-case analysis or pairwise deletion) to handling missing data (Allison 2002; Little and Rubin 2002). Provided certain assumptions about missingness are met, DML estimators maintain the usual asymptotic properties of ML estimators: consistency, asymptotic efficiency, and asymptotic normality (Allison 2002; Arbuckle 1996; Casella and Berger 2002; Little and Rubin 2002). MI also produces estimators that are consistent, asymptotically efficient, and asymptotically normal as the number of imputations approaches infinity (Allison 2002; Little and Rubin 2002; Schafer 1997). With respect to missing data, DML and MI estimators obtain these properties only under the assumption that the data are missing at random (MAR) or missing completely at random (MCAR) (Rubin 1976).

Data are MCAR if the probability of missingness is independent of the observed variables and independent of the values that are missing. This is unlikely to be true for the patterns of missingness sociologists typically encounter, but it is possible to test whether data are MCAR and proceed accordingly (Little and Rubin 2002). Data are MAR

[22] DML estimators are sometimes referred to as Full-Information Maximum Likelihood (FIML) estimators or Raw Maximum Likelihood (RML) estimators, particularly in the psychometrics literature (e.g., Arbuckle 1996).

if the probability of missingness is independent of the values that are missing conditional on the observed values. This is a less stringent criterion and is more likely to be encountered in sociological research. Unfortunately, it is not possible to test empirically whether data meet the MAR condition because any test would require precisely the information that is missing from the data.

The inability to test whether the MAR condition is met means that analysts must strive to include causes or correlates of missing values in their models. If a cause or correlate of missing data is left out of the model, then the data are missing not at random (MNAR) which, if treated using procedures for MAR data, can result in biased estimators. In some cases, however, variables related to missingness may be not of substantive or theoretical interest. Methodologists refer to such variables as auxiliary variables (AVs) and recommend that analysts identify and use AVs when addressing missing data (Collins, Schafer and Kam 2001; King et al. 2001; Rubin 1996; Schafer and Graham 2002; Schafer and Olsen 1998).

In their seminal study, Collins et al. (2001) defined two types of AVs. Type 1 AVs consist of variables associated with both a variable missing data and the pattern of missingness in that variable. Type 2 AVs consist of variables associated with a variable missing data, but not with the pattern of missingness. They found that under certain conditions, discussed in more detail below, Type 1 AVs can have an appreciable effect on reducing bias in an estimator and both Type 1 and Type 2 AVs can improve the efficiency of an estimator. Subsequent studies have generalized their results to measurement models and longitudinal data (Demirtas 2004; Yoo 2009), demonstrated how to include AVs in DML estimators (Graham 2003), and examined the performance

of AVs that are themselves missing data (Enders 2008). Some recent studies, however, have questioned the usefulness of AVs due to the stringent nature of the conditions in which they have been shown to reduce bias and the typically small efficiency gains. In simulation studies with arguably more realistic conditions for applied work, the benefits of AVs appear to be minimal (Mustillo forthcoming; Savalei and Bentler 2009; von Hippel 2007).

This study builds on prior studies in three ways to provide guidelines for sociologists looking to incorporate AVs into their handling of missing data with the specific goal of potentially reducing bias in an estimator. First, this paper offers a systematic examination of the properties an AV should satisfy in order to help an analyst meet the MAR condition and therefore reduce bias. This examination reveals that the properties are more strict than has been appreciated in past simulation studies and that some of the claims from the previous studies are inaccurate. Second, this paper presents the results from a simulation study that isolates and separately manipulates the associations between the AV and the variable missing data and between the AV and missingness. Past simulation studies have not isolated these two associations and therefore obscured potentially important dimensions in the performance of AVs. Third, this paper provides an empirical example involving AVs that illustrates a procedure for assessing candidate AVs.

What Are Auxiliary Variables

Methodologists define AVs as variables that are associated either with a variable missing data or the pattern of missingness in the variable and are not part of the model of interest. This definition contains two sources of ambiguity that deserve further

discussion. First, there is the idea that AVs may be *either* related to variables missing data *or* patterns of missingness. Methodologists generally consider two of the three possibilities implicit in this definition:[23] Type 1 AVs that are related to both the variable missing data and missingness in that variable and Type 2 AVs that are just related to the variable missing data and not to missingness in that variable (Collins, Schafer and Kam 2001). Type 1 AVs may help meet the MAR condition and therefore have the potential to reduce bias and improve the efficiency of an estimator. Type 2 AVs, however, cannot help meet the MAR condition, and therefore are only useful for improving the efficiency of an estimator. To see why this is the case, it is useful to consider a more formal treatment of AVs.

Suppose Y is an endogenous variable with some missing data, X is a covariate of interest, and Z is a candidate AV. We can define another variable R that is an indicator for missingness in Y that takes a value of 1 for cases missing values for Y and 0 otherwise. Further, suppose that in the absence of Z, the data are MNAR, that is

$$\Pr[R|Y, X] \neq \Pr[R|X]. \tag{4.1}$$

In other words, the conditional probability of missingness, R, depends on the endogenous variable, Y, even after conditioning on the covariate, X. If this is not the case and the data are already MAR without including Z, then the AV has no role to play in reducing bias in the estimator.

Similar to an argument made by Ibrahim, Lipsitz, and Horton (2001), if we consider the special case of linear associations, we can translate the conditional probability statement in (4.1) into a statement involving partial correlations as

[23] Methodologists have not examined the third type of AV, variables just related to the pattern of missingness in a variable.

$$\rho_{RY.X} \neq 0. \tag{4.2}$$

This equation simply states that a linear association between *R* and *Y* remains after controlling for *X* (i.e., a partial correlation between *R* and *Y*), which is indicative of an MNAR condition. If in addition to *X* we also control for *Z*, our candidate AV, then we have the second order partial correlation, $\rho_{RY.XZ}$, that gives the association between *R* and *Y* net of *X* and *Z*. The formula for calculating the second order partial correlation is given by (Stuart et al. 2005)

$$\rho_{RY.XZ} = \frac{\rho_{RY.X} - \rho_{RZ.X}\rho_{YZ.X}}{\sqrt{1-\rho_{RZ.X}^2}\sqrt{1-\rho_{YZ.X}^2}}. \tag{4.3}$$

From equation (4.3) we can see that the denominator is always non-negative and that the numerator will equal 0 only when the association between *R* and *Y* net of *X* is equal to the product of the associations between *R* and *Z* net of *X* and *Y* and *Z* net of *X* ($\rho_{RY.X} = \rho_{RZ.X}\rho_{YZ.X}$). Since by definition the association between *R* and *Y* net of *X* is nonzero ($\rho_{RY.X} \neq 0$), this expression can only move towards 0 when both the associations between *R* and *Z* net of *X* ($\rho_{RZ.X}$) and *Y* and *Z* net of *X* ($\rho_{YZ.X}$) are not equal to 0. This demonstrates in the linear case that in order for a candidate AV to help shift an MNAR to a MAR condition, the AV needs to be associated with both *Y* and *R* controlling for *X*.[24] The necessity of finding AVs that have associations with *Y* and *R* after adjusting for *X* has not been sufficiently emphasized in previous studies.

[24] Equation (3) also indicates that it is theoretically possible that including *Z* could lead to a higher partial correlation than just controlling for *X* (e.g., if $\rho_{RZ.X}$ and $\rho_{YZ.X}$ have opposite signs or in some cases when one of the associations is 0), which would indicate a potentially worse violation of the MAR condition and potentially greater bias. Such a situation has not been observed in simulation studies, but warrants further investigation.

The result that AVs need to be associated with both a variable missing data and with missingness in that variable in order to help meet the MAR condition is not surprising and is quite similar to the conditions known in the survey design literature for AVs used to form weights for survey nonresponse (Groves 2006; Kalton and Flores-Cervantes 2003; Kreuter and Olson 2011; Little and Vartivarian 2005). Some simulation studies, however, appear to demonstrate that using AVs that are only associated with the variable missing data and not missingness in the variable (i.e., Type 2 AVs) can reduce bias (Collins, Schafer and Kam 2001; Demirtas 2004). These results can be explained by a consideration of the data generating models underlying the simulations. For instance, in the Collins et al. (2001) study, data for the simulations were generated by first specifying a given association between Y and Z (e.g., $\rho_{YZ} = 0.9$) and then generating missingness in Y based on the values of Y itself. Because missing values for Y were generated without reference to Z, the authors claim that Z is not a cause of missingness. This is, of course, true in a strict sense, but misses the fact that there is an association between Z and R induced by the association between Z and Y. Even though Z was not a cause of missingness in the simulation, it does meet the criteria outlined above that it is associated with both Y and R, and therefore it is not surprising that it helps meet the MAR condition and reduces bias.

The second ambiguity in the definition of AVs is the idea that AVs are not part of the model of interest. We have just seen that AVs must be associated with Y to be beneficial in terms of reducing bias (and also for improving efficiency). This raises the question of when it is acceptable to have a variable associated with Y, especially an AV with a strong association with Y, that can be left out of the model of interest. It is likely

that AVs associated with *Y* will also share some association with the covariates and therefore leaving the AVs out of the analysis model can lead to omitted variable bias. There are at least two exceptions. First, analysts might be primarily interested in estimating descriptive statistics, in which case omitted variable bias is typically not a concern. Second, it may be possible to identify AVs that are endogenous and should not be included in the model. King et al. (2001) provide an example of this situation. Suppose one is interested in estimating the effect of partisan identification on voting behavior and one has information on how people intended to vote five minutes before voting. Voting intentions could serve as an AV if it is sufficiently associated with voting behavior and missingness in voting behavior, but in at least some analyses it would not be appropriate to include it as a covariate. Despite these exceptions, once candidate AVs are identified, it is important to consider whether in fact they belong in the substantive model and are not auxiliary to the analysis.

When Are Auxiliary Variables Beneficial

Studies analyzing the benefits of AVs have considered a number of different dimensions in their simulations. The three most important dimensions with respect to estimator bias include: (1) the strength of the association between AVs and a variable missing data, (2) the extent of missing data, and (3) the pattern of missingness in the data. The specific conditions for when AVs have an appreciable impact on reducing bias depend on a combination of all three of these dimensions and vary across different parameters.

Simulation studies have tended to examine two levels of association between AVs and variables missing data: (1) a moderate level with correlations ranging from 0.3

to 0.5 and (2) a high level with correlations ranging from 0.7 to 0.9 (Collins, Schafer and Kam 2001; Demirtas 2004; Enders 2008; Yoo 2009). A more recent study examined a range of associations between moderate and high correlations (Mustillo forthcoming). As a general rule, AVs with a moderate correlation with a variable missing data do not have much of an effect on bias, though there are some exceptions discussed below. More precisely, when systematic missingness largely depends on an AV with only a moderate correlation with a variable missing data, standard approaches to handling missing data based on a MAR assumption produce estimates with little bias and therefore AVs have little to contribute to reducing bias.

Simulation studies in this area typically introduce high rates of missing data, such as conditions with 25 percent missing, 50 percent missing, and sometimes even greater rates (Collins, Schafer and Kam 2001; Demirtas 2004; Enders 2008). Two studies have considered more modest rates of missing data ranging from 10 to 30 percent (Mustillo forthcoming; Yoo 2009). Not surprisingly, AVs have a larger impact on reducing bias at higher rates of missingness. In fact, AVs do not have much an effect on bias at rates of missingness up to 30 percent, except in a couple special cases. Again, this is primarily due to the fact that the AVs have little to contribute because standard approaches to handling missing data assuming MAR perform reasonably well.

Following Collins et al.'s (2001) lead, many of the simulation studies involving AVs have examined different missingness mechanisms. The mechanism underlying missingness is controlled by assigning missing values based on different functions of *Z*. The two most commonly used mechanisms involve linear and curvilinear functions. The linear mechanism is achieved by increasing the probability of assigning a missing value

in *Y* with increasing values of *Z*. Since *Y* and *Z* typically have a positive association, this results in higher values of *Y* having a higher chance to be missing. The curvilinear mechanism is achieved by setting high probabilities for missing in *Y* at the tails of *Z* and low probabilities for missing in *Y* at the center of the distribution for *Z*. The simulation studies indicate that linear patterns of missingness lead to more bias in estimates of means and intercepts, while curvilinear patterns of missingness lead to more bias in estimates of variances, regression coefficients, and residual variances in factor analysis models (Collins, Schafer and Kam 2001; Demirtas 2004; Yoo 2009). Furthermore, the missingness mechanism interacts with the extent of missingness and association between AVs and the variable missing data. For instance, with a curvilinear missingness mechanism, Collins et al. (2001) found that AVs with a moderate association with the variable missing data and in the 25 percent missing condition played an important role in reducing bias in a regression coefficient.

Simulation studies have identified the importance of the strength of the association between AVs and a variable missing data, the rate of missingness, and the mechanism underlying missingness in determining whether AVs are more or less likely to reduce bias in an estimator. These studies, however, have not isolated and manipulated the associations between the AV and the variable missing data and the AV and an indicator for missingness (*R*). As demonstrated in the previous section, both of these associations are important for reducing bias in an estimator. It may be that the strength of one of the two associations is more important or that both need to be relatively strong. It is this question that is the focus of the following simulation study.

Simulation Study

The following simulation study is designed (1) to analyze the relative importance of the associations between an outcome (*Y*) an AV (*Z*) and between missingness in *Y*, *R*, and *Z* in reducing bias and (2) to explore a mechanism for missingness, a threshold pattern, that has not been considered in previous studies of AVs. The study adopts a factorial design with four dimensions to accomplish these goals. The primary dimensions are the associations between the *Z*, an AV and *Y*, a variable missing data, and between the *Z* and *R*, an indicator for missingness in *Y*. This study examines a finer array of associations between *Z* and *Y*, 0.3 to 0.8 in increments of 0.1, than in past studies in order to provide more detailed information about the behavior of AVs. The specific associations between *Z* and *R* depend on the third dimension, the missingness mechanism.

This study analyzes two mechanisms for assigning missing values in *Y*. The first mechanism assigns missing values based on a curvilinear function (described in detail below). This mechanism was selected because it has been shown to be particularly problematic for bias in regression coefficients (Collins, Schafer and Kam 2001; Demirtas 2004; Yoo 2009) and to provide continuity with past studies. The second mechanism assigns missing values based on a threshold function. This mechanism has not been explored in studies of AVs, but it is a mechanism that can arise in practice, as will be illustrated in the empirical example following the simulation study. As Collins et al. (2001) note, the performance of AVs varies across different mechanisms for missingness, so it is valuable to analyze an approach to assigning missing values that has not been previously explored. The percentage of missing data is the final dimension.

This study considers two percentages, 25 percent and 50 percent, to maintain comparability with past studies.

Data Generating Model

This simulation study relies on the simple data-generating model used by Collins et al. (2001). Their model allows for an analysis of an estimated regression coefficient, the most common parameter estimated in sociological models. In addition, using the same data-generating model helps ensure comparability across studies.

Random samples of size $N = 500$ were drawn from a multivariate normal distribution for three variables, an outcome, *Y*, a covariate, *X*, and a candidate AV, *Z*. The distribution took the form

$$(Y, X, Z) \sim N(\mathbf{0}, \mathbf{\Sigma}), \tag{4.4}$$

with

$$\mathbf{\Sigma} = \begin{bmatrix} 1.44 & 0.60 & \phi \\ 0.60 & 1.00 & 0.6\phi \\ \phi & 0.6\phi & 1.00 \end{bmatrix}, \tag{4.5}$$

where the parameter ϕ ranged from 0.3 to 0.8 in increments of 0.1. Following Collins et al, (2001), the association between *Z* and *X* was set as a fraction of the association between *Y* and *Z* based on the association between *Y* and *X*. For instance, in the low correlation condition, ϕ is set to 0.3 and (4.5) becomes

$$\mathbf{\Sigma_{LC}} = \begin{bmatrix} 1.44 & 0.60 & 0.30 \\ 0.60 & 1.00 & 0.18 \\ 0.30 & 0.18 & 1.00 \end{bmatrix}. \tag{4.6}$$

Generating Missing Values

This study adopted two mechanisms for generating missing values in *Y*. The first mechanism closely approximates the nonlinear mechanism used in past studies (Collins, Schafer and Kam 2001; Demirtas 2004; Yoo 2009). This mechanism was implemented by constructing Z^2 and then constructing an indicator, *R*, for missingness in *Y* based on the following equations

$$R^* = Z^2 + \varepsilon \tag{4.7}$$

$$R = \begin{cases} 1 \text{ if } R^* \leq \tau \\ 0 \text{ if } R^* > \tau \end{cases}. \tag{4.8}$$

The variance of the error term in (4.7) was set to achieve specific associations between *R** and Z^2, and therefore indirectly between *R* and Z^2. The associations as measured by the coefficient of determination for *R** regressed on Z^2 ranged from 0.3 to 0.8 in increments of 0.1. The parameter τ was set to achieve the specific rates of missingness, 25 percent and 50 percent, based on the appropriate percentiles of the distribution of *R**.

The second mechanism for assigning missing values in *Y* was designed to replicate a threshold pattern. For this mechanism, an indicator for missingness was constructed based on the following equation

$$R = \begin{cases} 1 \text{ if } Z \leq \tau \\ 0 \text{ if } Z > \tau \end{cases}. \tag{4.9}$$

As with the first mechanism, the rate of missingness is set by choosing τ to achieve a 25 percent or 50 percent rates. With the threshold missingness mechanism, the choice of the rate of missing data dictates the association between *Z* and *R*. The association between *Z* and *R* is roughly -0.70 at 25 percent and roughly -0.80 at 50 percent.

Analysis

For the nonlinear missingness mechanism there are 72 cells in the factorial design (6 levels of association between *Z* and *Y*, 6 levels of association between *Z* and *R*, and 2 rates of missingness). For the threshold missingness mechanism there are 12 cells in the factorial design (6 levels of association between *Z* and *Y*, and 2 rates of missingness). For each of the 84 cells, Stata 12 was used to generate 10,000 samples of size $N = 500$ from the multivariate distribution in (4.4) with missing data created in *Y* as discussed above.

The simulation study relies on multiple imputation, the most common approach used in sociological research, to address missing data. Multiple imputation is an estimation strategy that consists of the following four steps: (1) replication, (2) imputation, (3) analysis, (4) recombination (Little and Rubin 2002). The first step, replication, involves making *m* copies of the incomplete data set. The second step, imputation, involves replacing missing values in each incomplete data set with imputed values including a random component and thus forming *m* complete data sets. The third step, analysis, involves estimating the desired model on each of the complete data sets. The final step, recombination, involves using formulas developed by Rubin (1987) to combine the results across the data sets to obtain parameter estimates and standard errors that appropriately reflect the additional uncertainty from the missing data.

For every sample multiple imputation was invoked twice to create 10 complete data sets each time. The first round of multiple imputation only included *X* in the imputation model. This serves as a point of comparison while the second round included *X* and *Z*, the auxiliary variable. The imputation model is based on a linear

regression model. Let y_i be partitioned in vectors containing complete and incomplete observations (y_{io}, y_{im}), and similarly partition $\mathbf{x} = (\mathbf{x}'_o, \mathbf{x}'_m)$, which consists of x_i or x_i and z_i depending on the round of multiple imputation. The steps in imputation involve first fitting a regression model for y_{io} to the observed data to obtain estimates of $\hat{\beta}$ and $\hat{\sigma}^2$. Second, new parameters β_* and σ_*^2 are drawn from the joint posterior distribution with a noninformative improper prior $\Pr(\beta, \sigma^2) \propto 1/\sigma^2$ (StataCorp 2011). Third, a set of imputed values is obtained for y_{im} by drawing from $MVN(\mathbf{X}\beta_*, \sigma_*^2\mathbf{I})$. Finally, the second and third steps are repeated 9 more times to obtain 10 complete data sets.

For each sample, regardless of whether *Z* was included in the imputation model, *Y* was regressed on *X*, the results were combined across the completed data sets using Stata's suite of multiple imputation commands, and the estimated regression coefficients were stored for further analysis. In the case of a bivariate regression, the regression coefficient is simply the ratio of the covariance between *Y* and *X* divided by the variance of *X*. Given the covariance matrix in (4.6), the population regression coefficient is 0.6. Bias for the multiple imputation estimators that did and did not include the AV was calculated as the difference between the average estimate for the regression coefficient and the population value within each cell of the factorial design. In addition, Collins et al. (2001) recommend a standardized measure of bias calculated as

$$Std.\,Bias = \frac{E[\hat{\beta}]-\beta}{SD[\hat{\beta}]}x100 \qquad (4.10)$$

to facilitate interpretation.[25] They suggest that a value of standardized bias greater than 40 represents a problematic level of bias for empirical analyses in social science research. Cells with standardized bias values greater than 40 are indicated with asterisks in the tables of results.

Simulation Results

Table 11 presents the results from the simulations involving the nonlinear missingness mechanism. Under all conditions, the multiple imputation estimator for the regression coefficient appears to be unbiased when the AV is included in the imputation model (see Panel B). This result is consistent with past studies, particularly Collins et al. (2001). In addition, there is some degree of bias in the multiple imputation estimator of the regression coefficient under all conditions when the AV is not included in the imputation model (see Panel A). The extent of the bias when the AV is not included varies across all of the dimensions of the simulation. The bias increases with more missing data, as the association between *Y* and *Z* increases, and as the association between *R* and *Z* increases. Based on the Collins et al. (2001) threshold of a value of 40 for the standardized bias, the extent of bias is generally not substantively problematic until the correlation between *Y* and *Z* reaches 0.6 in the 25 percent missing condition. In the 50 percent missing condition, the extent of bias is not problematic until the correlation between *Y* and *Z* reaches 0.5 or the correlation between *Y* and *Z* reaches 0.4 coupled with an R^2 from regressing R^* on Z^2 of 0.6 or higher. The general lower right triangular pattern of problematic bias illustrates that both the association between *Y*

[25] Simulation studies that examine bias often rely on a measure of percent bias. This study reports a measure of standardized bias to maintain comparability with past simulations involving AVs.

and *Z* and the association between *R* and *Z* are important when assessing the potential benefit of an AV.

Table 12 presents the results from the simulations involving the threshold missingness mechanism. The multiple imputation estimator shows virtually no bias across both dimensions, the association between *Y* and *Z* and the rate of missing data, when the AV is included in the imputation model (see Panel B). In contrast, there is some degree of bias in all conditions when the AV is not included in multiple imputation model. The bias reaches the threshold for substantive importance once the correlation between *Y* and *Z* reaches 0.4 for both the 25 percent and 50 percent conditions and increases as the correlation strengthens.

This simulation study was designed to analyze the relative contributions of the associations between *Y* and *Z* and between *R* and *Z* as well as the threshold missingness mechanism. The results from the nonlinear missingness mechanism indicate, as anticipated, that both associations play an important role in determining whether or not including AVs eliminates substantial estimator bias. An AV that has a weak association with either *Y* needs to have a particularly strong association with *R*, and vice versa, in order to lead to a substantial reduction in bias. The results from the threshold mechanism did not follow the same pattern. The rate of missing data dictated the strength of the association between *Z* and *R*. The extent of bias was slightly larger in the 50 percent missing condition, but whether the bias was substantively problematic did not vary across the two conditions. This is most likely due to the relatively high associations between *R* and *Z* in both conditions. Finally, the difference in the pattern and magnitude of bias across the two missingness mechanisms underscores a point

Table 11: Bias in Regression Coefficient from Nonlinear Missingness Simulations.

	No AV Used in Imputation Model					
	25% Missing					
Y,Z \ R,Z	0.3	0.4	0.5	0.6	0.7	0.8
0.3	0.005	0.005	0.006	0.007	0.007	0.008
0.4	0.008	0.010	0.011	0.012	0.012	0.015
0.5	0.013	0.015	0.017	0.019	0.021	0.023*
0.6	0.018	0.022*	0.024*	0.028*	0.030*	0.032*
0.7	0.024*	0.029*	0.032*	0.037*	0.040*	0.044*
0.8	0.032*	0.037*	0.042*	0.048*	0.052*	0.056*
	50% Missing					
	0.3	0.4	0.5	0.6	0.7	0.8
0.3	0.010	0.012	0.015	0.018	0.019	0.023
0.4	0.018	0.022	0.027	0.030*	0.034*	0.039*
0.5	0.029*	0.034*	0.040*	0.046*	0.053*	0.058*
0.6	0.040*	0.049*	0.058*	0.066*	0.074*	0.081*
0.7	0.055*	0.067*	0.08*	0.088*	0.098*	0.107*
0.8	0.068*	0.084*	0.100*	0.113*	0.124*	0.136*
	AV Used in Imputation Model					
	25% Missing					
Y,Z \ R,Z	0.3	0.4	0.5	0.6	0.7	0.8
0.3	0.000	0.000	0.000	0.000	0.000	0.000
0.4	0.000	0.001	0.000	0.000	-0.001	0.001
0.5	0.001	0.000	0.000	0.000	0.000	0.000
0.6	0.000	0.001	0.000	0.000	0.001	0.000
0.7	0.000	0.001	-0.001	0.000	0.000	0.000
0.8	0.001	0.000	0.000	0.001	0.001	0.000
	50% Missing					
	0.3	0.4	0.5	0.6	0.7	0.8
0.3	0.000	-0.001	0.000	0.000	-0.001	0.001
0.4	0.000	-0.001	0.000	-0.001	0.000	0.001
0.5	0.000	-0.001	-0.001	0.000	0.001	0.000
0.6	0.000	0.000	0.000	0.000	0.000	0.000
0.7	0.001	0.000	0.002	0.000	0.000	-0.001
0.8	0.000	-0.001	0.000	0.001	0.000	0.000

Notes: Association between *Y* and *Z* expressed as the correlation.
Association between *R* and *Z* expressed as the R^2 from regressing *R** on Z^2.

first made by Collins et al. (2001) that impact of missing data on an analysis can significantly depend on the mechanism(s) underlying the missingness.

Table 12: Bias in Regression Coefficient from Threshold Missingness Simulation.

	No AV Used in Imputation Model	
	% Missing	
Y,Z	25%	50%
0.3	-0.016	-0.022
0.4	-0.030*	-0.040*
0.5	-0.047*	-0.063*
0.6	-0.068*	-0.096*
0.7	-0.094*	-0.135*
0.8	-0.128*	-0.181*
	AV Used in Imputation Model	
	% Missing	
Y,Z	0.3	0.5
0.3	0.000	0.000
0.4	0.000	0.000
0.5	0.000	0.002
0.6	0.000	0.000
0.7	0.000	0.000
0.8	-0.001	0.002

Notes: * indicate cells with std. bias > 40. Association between *Y* and *Z* expressed as the correlation.

Identification and Evaluation of AVs in Practice

Past work on AVs has largely focused on the results of simulation studies with little consideration of how AVs could be identified and assessed in empirical research (though, see Baraldi and Enders (2010) and Mustillo (forthcoming) for exceptions). This section provides an extended consideration of the assessment of AVs in an empirical setting with a focus on identifying and assessing the potential benefits of candidate AVs. The example comes from one component of a larger study analyzing the effects of the fetal environment on adult SES, BMI, and hypertension. The specific

component involves developing a measurement model for the fetal environment (see Bollen, Adair, and Noble (2011) for an extended discussion of this analysis). The measurement model is a two-factor confirmatory factor analysis (CFA) with latent variables for birth weight and gestational age, two aspects of the fetal environment.

Data for this example come from the Cebu Longitudinal Health and Nutrition Survey, a community-based study involving more than 3000 newborns in Cebu, Philippines (Adair et al. 2011). The study included two measures of birth weight and two measures of gestational age (see Table 13 for descriptive statistics). The two measures of birth weight consist of one measure taken at the place of delivery and another measure taken by project staff. One measure of gestational age is based on the mother's self-report of the number of weeks since her last menstrual period. The other measure is a Ballard score, a score derived from a clinical assessment based on six developmentally staged neuromuscular characteristics and six physical infant characteristics (Ballard, Novak and Driver 1979).

The challenge with this analysis is that the Ballard is missing data for over 80 percent of the newborns. When faced with this much missing data, the usual response would be to remove the variable from consideration. There are, however, two factors that suggest otherwise. First, the other measure of gestational age, the self-reported last menstrual period, is known to be highly unreliably. Second, the reason the Ballard is missing for so many newborns is that they had to meet at least one of a set of conditions in order for it to be administered. The conditions for administering the Ballard included: (1) the infant was born weighing less than 2500 g, (2) the mother reported bleeding during her pregnancy, (3) the mother can't remember her last menstrual

period, (4) the mother did not have a menstrual period between the birth of a previous child and conception of the index child, and (5) the mother had diabetes during pregnancy. These conditions are available for most of the newborns in the data and provide information about the Ballard.

Table 13: Descriptive Statistics.

	Valid N	% Miss	Mean	SD
Birth weight 1	2555	14%	30.31	4.74
Birth weight 2	2958	1%	29.95	4.35
Last menstrual period	2773	7%	36.62	0.74
Ballard	579	81%	36.36	0.48
Low birth weight	2981	0%	0.10	0.30
Bleeding	2986	0%	0.07	0.25
Can't recall LMP	2986	0%	0.07	0.26
No mentrual period	2371	21%	0.07	0.26
Diabetes	2960	1%	0.01	0.08

Notes: Gestational age measures are logged weeks x 10. Birth weight measures are 100 g. Low birth weight is defined as less than 2,500 g. LMP stands for last menstrual period.

The conditions for administering the Ballard are excellent candidates for AVs. These variables clearly relate to missingness on the Ballard and may also relate to the values of the Ballard itself. In addition, the conditions account for virtually all of the missing data for the Ballard, and therefore the mechanism underlying missingness is known. Knowledge of the mechanism helps with specifying the functional form for the relationship between the AVs and missingness. Finally, the AVs are not of substantive interest in the model.[26]

[26] The first condition related to low birth weight is, of course, a function of birth weight, which is a variable of substantive interest. It as an empirical question, addressed below, whether low birth weight provides additional information about missingness on the Ballard beyond birth weight itself.

Assessment of Candidate AVs

Based on past results and results in this article, one can outline a number of criteria to assess whether candidate AVs are more or less likely to be significantly beneficial in reducing bias. The first is the extent of missing data. Simulation studies suggest that if a target variable is missing for less than 25 percent of cases, an AV is unlikely to be beneficial for reducing bias (Mustillo forthcoming). With this example, the Ballard is missing for over 80 percent of cases, so the rate of missingness is clearly high enough for AVs to have potential benefits in reducing bias.

The second thing to check is the strength of the association between the candidate AVs and the variable missing data. This study and past simulation studies show that the threshold level of association for AVs to beneficial in reducing bias varies across different mechanisms for missingness, different parameters, and the strength of the association between the AVs and missingness (Collins, Schafer and Kam 2001; Demirtas 2004; Enders 2010; Mustillo forthcoming; Yoo 2009). However, the minimum association between an AV and a variable missing data to see benefits in terms of reducing bias appears to be around a correlation of 0.4. This translates into an R^2 of 0.16 if considering the association between multiple AVs and a variable missing data. As discussed above, this association should be net of the substantive variables in the model.

In practice, it is impossible to directly assess the strength of the association between AVs and the variable missing data. The quantity one is most interested in is the association between, say, $\mathbf{Z}$ and Y, but what one can calculate is the association between $\mathbf{Z}_{obs}$ and Y_{obs}. It is in general unknown how well the association between $\mathbf{Z}_{obs}$ and Y_{obs}

reflects the true association between **Z** and *Y*. Nonetheless, it can still be informative to examine the association between **Z** and *Y*. In this example, there are multiple candidate AVs, so the R^2 is a useful measure of association. Two models are estimated to isolate the association between the Ballard and the candidate AVs net of substantive variables of interest.[27] The first model regresses the Ballard on all of the substantive variables of interest and the candidate AVs. The second model regresses the Ballard on just the substantive variables. The difference in the R^2s from the two models serves as a measure of association between the Ballard and the AVs net of the substantive variables.

Table 14 presents the R^2s from these regressions and the difference. Although the substantive variables explain about 20 percent of the variance in the Ballard, the candidate AVs account for virtually none of it. This is not too surprising given that the candidate AV that is most likely related to the Ballard from a theoretical standpoint, low birth weight, is already represented in the substantive model. It is possible, however, that the association between the Ballard and the AVs among the cases with observed data does not reflect the underlying association that would be observed with complete data, particularly given the extent of missing data.

[27] In this example, these models are estimated on the complete case data. It may be possible to develop a procedure that would allow one to assess the AVs on the full sample, but this adds a number of additional complications and remains for future research.

Table 14: Strength of Association Between AVs, Ballard, and Missingness.

	Ballard	Miss Ballard
R^2 model 1	0.203	0.647
R^2 model 2	0.201	0.140
Difference	0.002	0.507

Notes: Model 1 includes the substantive variables and the candidate AVs. Model 2 just includes the substantive variables. The difference provides the strength of association between the Ballard the AVs net of the substantive variables. The models for Missing Ballard are linear probability models.

The third quantity to check is the association between the AVs and an indicator for missing data on the Ballard. The simulation reported above is the first to isolate and manipulate this association, so there is less information to develop guidelines for what to expect. The simulation results suggest that the strength of the association required for AVs to have a significant effect on bias varies across missingness mechanisms and according to the strength of the association between the variable missing data and the AVs. In addition, 0.4 appears to be a reasonable threshold for the minimum correlation between an AV and missingness in a variable, though more research is needed across a wider variety of contexts to offer a more definitive assessment.

There are a couple options for assessing the strength of the association between a set of candidate AVs and an indicator for missingness in the target variable. One could estimate a linear probability model or a logit (or probit) model. The linear probability model has the advantage that one can use the same procedure involving R^2s to isolate the effect of the candidate AVs net of the substantive variables. Although the R^2 from a linear probability model is known to not to be strictly less than the R^2 from an analogous linear regression model (Aldrich and Nelson 1984), it still serves as a rough indication of the strength of association. Table 4 presents the R^2s from estimating linear

probability models regressing an indicator for missing data on the Ballard on the candidate AVs and the substantive variables. As expected, the candidate AVs explain a substantial portion of the variance in the indicator and appear to be well above the threshold for seeing potential benefits in terms of reducing bias.

Parameter Estimates from Substantive Model

The candidate AVs in this example meet two out of three criteria for seeing beneficial effects in reducing bias. There is a high rate of missing data in the Ballard (over 80 percent) and there is a strong association between the AVs and missingness on the Ballard. The AVs, however, are not associated with the Ballard itself, at least among the observed values of the Ballard. Next, we compare the parameter estimates from the substantive model with and without using the AVs.

The missing data was addressed using multiple imputation with chained equations as implemented in Stata 12 (StataCorp 2011). The chained equations approach addresses the missing data in the other covariates as well as the Ballard. Two rounds of multiple imputation were run to construct 10 complete data sets each time. The first round did not include the AVs in the imputation model, the second round did include the AVs. The measurement model, a two-factor CFA with two indicators for birth weight and two indicators for gestational age, was estimated in Mplus 6 for each set of multiply imputed data sets (Muthén and Muthén 1998-2010).

Table 15: Parameter Estimates for Models With and Without AVs.

	No Avs		Avs		
	Est	SE	Est	SE	% Diff.
L BW1	0.931	(0.081)	0.943	(0.060)	-1%
L Ballard	1.163	(0.138)	1.044	(0.169)	10%
V(BW)	16.846	(1.593)	16.695	(1.148)	1%
V(GA)	0.046	(0.013)	0.045	(0.016)	2%
Cov(BW, GA)	0.713	(0.066)	0.739	(0.063)	-4%
V(BW1)	7.838	(1.297)	7.721	(0.927)	1%
V(BW2)	2.488	(1.504)	2.740	(1.083)	-10%
V(LMP)	0.503	(0.018)	0.508	(0.020)	-1%
V(Ballard)	0.156	(0.015)	0.174	(0.017)	-12%

Notes: BW refers to latent birth weight. GA refers to latent gestational age. BW1 and BW2 refer to the first and second measures of birth weight. LMP refers to the last menstrual period.

Parameters from the measurement model include the factor loadings for one of the birth weight measures and for the Ballard,[28] the variances of the latent variables, the covariance between the two latent variables, and the residual variances for all of the indicators. There is little substantive difference in most of the parameter estimates obtained from the multiply imputed data with and without the AVs (see Table 15). In percentage terms, the largest difference in estimates is for the residual variance of the Ballard. The estimate based on the imputed data with no AVs is 12 percent less than the estimate based on the imputed data using AVs. The second largest difference is in estimate of the factor loading for the Ballard, which is 10 percent larger when using the imputed data with no AVs. Finally, although it is not the focus of this work, it is interesting that there appears to be some substantial efficiency gains from using the

[28] The factor loadings for the other birth weight measure and the last menstrual period were constrained to 1 to provide a scale for the latent variables and identify the model.

AVs. This is surprising given the lack of an association between the AVs and the Ballard (at least for the observed cases) and warrants further investigation in future research.

This example was included to illustrate an approach to assessing candidate AVs and to evaluate the performance of AVs in an empirical setting. The conditions for administering the Ballard appeared to be good candidate AVs; however, the lack of an association, in particular, net of birth weight, between the AVs and the Ballard suggested that they may not have much effect on reducing potential bias (i.e., these would be Type 3 AVs). Although it is not possible to know the true parameter values for this analysis, the fact that the estimates do not significantly differ when AVs are or are not included in the imputation model indicates that the AVs do not have much effect on reducing potential bias.

Conclusion

Methodologists have recently begun to recommend the routine incorporation of AVs into procedures for handling missing data. These recommendations are based on the results of simulation studies that demonstrate that in some circumstances AVs can reduce bias (and improve efficiency) and that including AVs that are not related to variables missing data does not appear to increase bias (or decrease efficiency). The goal of this research was to provide more information for sociologists looking to use AVs when addressing missing data in their analyses and to recommend that practitioners assess the potential benefits of AVs before simply including them in their handling of missing data.

A more formal treatment of AVs using Rubin's framework for missing data and considering the case of linear associations clarified that in order for AVs to reduce bias

by shifting an MNAR setting closer to a MAR setting they need to be associated both with a variable missing data and an indicator for missingness on that variable. This dual condition has not been sufficiently emphasized in past studies of AVs, though it has clear implications for assessing candidate AVs. Even though it is not possible to know for certain the underlying relationship between a set of AVs and a variable with missing data, the relationship among the cases with observed values can be informative. It is easier to assess the relationship between the AVs and an indicator for missingness in a variable.

The simulation study provided more information about the strength of the associations between an AV and a variable missing data and an AV and missingness. The results of the simulation replicated past findings that a correlation around 0.4 between an AV and a variable missing data is a minimum for seeing beneficial effects from using the AV. Moving beyond past studies, the results indicated that a similar minimum correlation is needed between an AV and missingness. In addition, the simulation analyzed a threshold missingness mechanism, a mechanism that has not been explored in previous studies, and found that minimum correlation of 0.4 between the AV and a variable missing data holds for this mechanism as well.

Finally, an empirical example illustrated how one can assess the likely benefits of including AVs in the handling of missing data. The example showed that even seemingly good candidate AVs may not make much of a difference in the parameter estimates. With this example, the reason the AVs did not have much an effect is likely due to a low association between the AVs and the variable missing data.

This work, consistent with a couple of recent studies, suggests that in practice it may be difficult to find AVs that will make much of a difference for an analysis. Although it appears that AVs are unlikely to increase bias and therefore there is little harm in including them when addressing data, it is still useful to know whether they are likely to be beneficial.

CHAPTER V: CONCLUSIONS

This dissertation aimed to make two broad contributions. First, the studies in chapters 2 and 3 contribute to our understanding of the implications of the expansion of higher education over the course of the 20th century. The increase in college enrollment has had two effects: (1) some people who currently attain college degrees would most likely not have in the past and (2) the expansion has been disproportionately concentrated among lower tiered colleges and universities. These two considerations suggest that there may be more variation in the quality of higher education and, as a consequence, the returns to a college degree among recent cohorts than past cohorts. Recent studies have begun to examine the variation in the returns to higher education for earnings, civic participation, and fertility. Chapters 2 and 3 extend this line of work to measures of health and wellbeing. The beneficial effect of a college degree on health is central to our understanding of health inequalities. Thus a more complete sense of variation in the beneficial effects of higher education contributes to both research on the expansion of higher education as well as research on health inequalities.

Second, the methodological work in chapter 4 was initiated as a means of determining whether auxiliary variables would be beneficial in addressing the missing data in the analyses of the health-related returns to a college degree. The results of this study informed the handling of missing data in the earlier chapters in the dissertation,

but they also contribute more generally to our understanding of the conditions under which auxiliary variables may be beneficial quantitative sociological research. In addition, the chapter includes a discussion and illustration of a procedure for assessing auxiliary variables in practice.

The remainder of this conclusion is organized as follows. First, the key results from each of the chapters are reviewed. Second, consideration is given to a comparison of the conclusions across chapters 2 and 3, specifically with respect to the implications for our understanding of expansion of higher education. The final section discusses some of the limitations of the analyses in this dissertation and directions for future research.

Variation in the Benefits of a College Degree for Health

Chapter 2 focused on an analysis of variation in the returns to a college degree for two health behaviors (smoking and BMI) and two health outcomes (self-rated health and systolic blood pressure) among a recent cohort of young adults. In addition, the analytic approach allowed for an assessment of two competing theories of the role of education in contributing to or alleviating health inequalities: resource substitution theory and resource multiplication theory. Resource substitution theory posits that higher education compensates for a deficit of other resources that can promote better health, and therefore education serves to reduce health-related inequalities. In contrast, resource multiplication theory posits that higher education magnifies the benefits of other resources that promote health. To the extent this theory holds, higher education contributes to health-related inequalities.

Given that past work has identified that the health-related benefits of a college degree are unevenly distributed across males and females and across racial and ethnic groups, all analyses were conducted separately for black and white, women and men. The results suggest that there is variation in the returns to a college degree among white females for self-rated health and among white males for self-rated health and BMI. The particular pattern of variation is consistent with resource multiplication theory. White women and men who attain a college degree are estimated to accrue more health-related benefits from their degree than would be the case for white women and men who did not attain a college degree were they to do so.

The results suggest that there is also variation in the health-related returns to a college degree for black females. Higher education is associated with better self-rated health, lower systolic blood pressure, and a reduced likelihood of smoking among black women. Furthermore, there is evidence of variation in the returns to a college degree for self-rated health and smoking that is consistent with resource substitution theory. Among black women, the estimated health-related benefits of higher education are greater for people who did not attain a degree were they to do so than those who did attain a degree.

Finally, the results for black males are inconclusive. The diagnostics from the construction of the propensity score weights indicated that it is difficult to construct weights that balance the covariates for those who did and did not attain a degree. In particular, it is hard to identify black males who did not attain a college degree but had a high propensity to do so. This is most likely an indication that there are unobserved factors that differentiate black males who do and not attain college degrees. These

results indicate not only that it is difficult to identify variation in the health-relate returns to higher education among black males, but that even the average estimate of the benefits of higher education is likely to be biased.

Taken as a whole, the analysis in Chapter 2 provides evidence of variation in the health-related returns to a college degree that goes beyond what has been previously uncovered. Within-group analyses conducted for a recent cohort of young adult black and white, females and males found variation for three out of the four groups and inconclusive results for the fourth group. The particular pattern of variation was not constant across the groups, but rather suggests that higher education may play a different role with respect to other resources for whites than for black females.

Variation in the Protective Effect of a College Degree for Depression

Whereas chapter 2 focused on measures of physical health, chapter 3 addressed psychological wellbeing. The analysis of variation in the protective effects of higher education for depression proceeded in much the same way as in chapter 2, except that it also required addressing measurement error in the indicators of depressive symptoms (a subset of the CES-D scale). Also, as in chapter 2, the analytic approach allowed for an assessment of the relative importance of resource substitution theory and resource multiplication theory.

Past work has found little evidence of a main protective effect of education in general for depression among blacks. The analysis in chapter 3, however, shows that a college degree does have a protective effect against depression for black females. There are a number of possible explanations for the difference in results – the current study addresses measurement error in the indicators of depressive symptoms, while most

past studies do not; the current study focuses on depression among a recent cohort of young adults, while most past studies include a broader range of ages and cohorts; the current study focuses on higher education, while most past studies use either years of education or a broader range of educational degrees; the current study conducts separate analyses for males and females, while most past studies combine the sexes – thus more work is needed to determine the source of the discrepant results. Nonetheless, the finding in this analysis suggests that higher education may have more benefits for black women with respect to depression than has been appreciated based on past work.

In addition, there is evidence of variation in the protective effect of a college degree on depression among black women. The particular pattern of variation is consistent with resource substitution theory. Black women who did not earn a college degree are estimated to derive a greater protective from a degree were they to earn one than black women who did earn a college degree.

As noted in the analysis in chapter 2, the results for black men were inconclusive due to the challenge in finding black men with and without college degrees that were sufficiently comparable to warrant estimating the protective effect of a college degree. Although it is not possible to determine whether the likely presence of unobserved factors that influence both attaining a college degree and depression has led to upwardly or downwardly biased estimates, the results of this analysis suggest we should be cautious in interpreting past estimates as causal effects. In this case, it may be that a college degree has a protective effect on depression for black males, but more work is needed to find this out.

Among white women and men there is little evidence of variation in the protective effects of higher education on depression. As expected based on past work, a college degree protects against depression for white women. In contrast, a college degree appears to have no effect on depression for white males.

Auxiliary Variables in Sociological Research

Chapter 4 focuses on an analysis of the use of auxiliary variables (AVs) when handling missing data. Methodologists recommend incorporating AVs into procedures for missing data in order to reduce potential bias by helping to meet the missing at random (MAR) condition. Given the extent of missing data in some of the measures of family SES, AVs had the potential to be useful in the analyses of variation in the health-related returns to education.

The analytic results based on Rubin's framework for missing data clarified that in order for AVs to reduce bias by shifting an MNAR setting to a MAR setting they need to be associated with both a variable missing data (Y) and the pattern of missingness in that variable (R). The need for both associations has not been sufficiently appreciated in past studies that recommend the routine inclusion of AVs. Although it is possible in theory to find variables that predict both Y and R, it was not possible to find such variables for the analyses in chapters 2 and 3.

The simulation study provided additional evidence about the strength of the associations AVs need to have with Y and R to be beneficial in terms of reducing bias. Across two different mechanisms for assigning missing data, the associations between an AV and Y and between an AV and R needed to be quite substantial, at least a correlation of 0.4, in order for the AV to have an appreciable effect on estimator bias.

Thus, not only do AVs need to be associated with *Y* and *R*, they also need to have non-trivial relationships.

Finally, an empirical example illustrated how one can assess the likely benefits of including AVs in the handling of missing data. Given that it may be difficult to find reasonable candidates for AVs in a practice, as was the case for the analyses in chapters 2 and 3, it is important to assess the potential benefits of AVs rather than simply including them in the handling of missing data.

General Conclusions

One of the main goals of this dissertation was to develop a better understanding of the potential ramifications of the expansion of higher education during the 20th century. For the reasons outlined above, the expansion has likely led to an increase of the variation in the returns to a college degree among recent cohorts. The results from chapters two and three provide evidence of meaningful variation in the health-related benefits of higher education among white women and men and black women for important indicators of physical health and among black women for psychological wellbeing. Until recently, most research has implicitly assumed that people realize essentially the same returns to a college degree (or, more precisely, that any variation in the returns was random or due to a handful of individual attributes, such as sex). This has led to calls for policies that would allow everyone to attain a college education. A more nuanced understanding of variation in the returns to a college degree suggests that the estimated benefits from past studies may be misleading.

Another goal of this dissertation was to develop a better understanding of how higher education relates to health inequalities. With respect to this question, the results

provide a more complex picture. For black women, there is evidence that a college degree could contribute to reducing within-group health inequalities with respect to physical health and depression. As such, policies that promote access to higher education may be especially beneficial, as higher education appears to compensate for other resource disadvantaged among black females. In contrast, for white women and men, a college degree appears to magnify the physical health-related benefits of other background resources, and therefore contributes to within-group health inequalities. As discussed in the next section, more research needs to be done to identify the mechanisms that underlie these patterns of results.

Finally, an unanticipated, but robust, finding from chapters 2 and 3 was the difficulty in obtaining reliable estimates of the health-related returns to a college degree for black men. Although Add Health data contains a rich set of controls that includes measures of the most likely predictors of both educational attainment and the health outcomes (e.g., measures of parent SES, measures of ability, …), these covariates were not sufficient to adjust for preexisting differences between black men who attain a college degree and black men who do not. This suggests that there are other unobserved characteristics that can account for these differences. More work is needed to identify what the unobserved factors that affect educational attainment and health and wellbeing may be and to incorporate them into analyses in order to obtain more reliable estimates of the health-related returns to a college degree for black men.

Limitations and Future Research

There are number of directions to extend research on variation in the health-related benefits of higher education that come out this dissertation. First, the analytic

strategy was used to identify the presence of heterogeneity in the returns to a college degree without having to rely on interaction terms. This approach avoids the "curse of dimensionality" that arises with the number of possible interactions. The limitation in this analytic approach, however, is that it does not allow for direct tests of precise mechanisms that could explain the sources of the variation in the effects of higher education. Future work should work towards identifying the sources of variation. Based on the pattern of results with respect to resource substitution theory and resource multiplication theory, different measures of background resources would be good candidates to begin with.

Second, the analyses in chapters 2 and 3 drew on Add Health data. At Wave 4 the respondents in Add Health are in their late 20s and early 30s. For these analyses it was beneficial to work with a limited age range, but future work should look to assessing variation in the health-related benefits of higher education at other points in the life course. Past work has shown that the effects of education increase up to a threshold in old age (though this remains an active area of research). If true, it may be possible to detect more variation in the health-related returns to a higher degree among older respondents. In addition, although Add Health respondents show some signs of deteriorating health, older respondents are more likely to have significant health problems.

Finally, a central idea that orients this dissertation is that the expansion of higher education has led to an increase in the variation of the returns to a college degree. This is, of course, a comparative statement between past and recent cohorts. The analyses in chapters 2 and 3 focused on a sample of people from a recent cohort,

and therefore cannot directly speak to whether there is more variation in the health-benefits of higher education among recent cohorts than past cohorts. Future work should look to compare the extent of variation in the returns to education across cohorts, though this will be challenging on at least three fronts. First, it is not clear whether the analytic strategy used in this dissertation can easily be extended to an analysis that accounts for cohort effects. Second, simply decomposing the average effect of education on health into age, period, and cohort components remains a difficult area of research. Third, the data requirements are stringent and it is difficult to find sources of data that would allow for such an analysis. Nonetheless, as this research and other studies document the extent of variation in the returns to a college degree across a range of outcomes, it will ultimately be important to consider change over time.

APPENDIX ONE: DESCRIPTIVE STATISTICS FOR COVARIATES (HEALTH)

Table 16: Descriptive Statistics for Covariates of Health.

	Mean
Wave 4 Age	28.50
Region: West	0.14
Region: Midwest	0.31
Region: South	0.42
Region: Northeast	0.13
HH: 2 bio parents	0.51
HH: 2 parents	0.19
HH: single mother	0.22
HH: single father	0.03
HH: other	0.05
Mother's education	5.68
Father's education	5.58
Log family income	3.57
Lived environment	3.37
PVT score	102.18
GPA	2.80
Birth weight	117.13
Wave 1 SRH	3.90
Wave 1 BMI	22.57
Wave 1 daily smoke	0.01
Wave 1 physical activity	2.08
Wave 1 drinking	2.09

Note: Based on 10 complete data sets. Weighted data.

APPENDIX TWO: PROPENSITY SCORE MODELS (HEALTH)

Table 17: Propensity Score Models for Health Outcomes Analysis.

	Black Female			Black Male		
	Est	SE		Est	SE	
Wave 4 Age	-0.074	(0.054)		-0.037	(0.068)	
Region: Midwest	-0.364	(0.361)		0.383	(0.455)	
Region: South	0.055	(0.304)		0.107	(0.396)	
Region: Northeast	0.069	(0.436)		0.425	(0.618)	
HH: 2 parents	-0.699	(0.278)	*	-1.001	(0.372)	**
HH: single mother	-0.386	(0.241)		-0.521	(0.310)	
HH: single father	-0.564	(0.524)		0.704	(0.832)	
HH: other	-1.093	(0.380)	**	0.129	(0.427)	
Mother's education	0.147	(0.240)		-0.651	(0.330)	*
Mother's education^2	0.004	(0.022)		0.075	(0.029)	**
Father's education	-0.180	(0.270)		0.089	(0.324)	
Father's education^2	0.028	(0.025)		0.008	(0.030)	
Log family income	0.528	(0.687)		-0.199	(0.713)	
Log family income^2	-0.031	(0.108)		0.051	(0.123)	
Lived environment	0.091	(0.152)		0.118	(0.230)	
PVT score	0.048	(0.009)	***	0.029	(0.010)	**
GPA	1.107	(0.161)	***	1.202	(0.188)	***
Birth weight	0.006	(0.005)		0.001	(0.007)	
W1 SRH	0.249	(0.101)	**	0.426	(0.144)	**
W1 BMI	0.007	(0.019)		0.032	(0.027)	
W1 daily smoke	0.622	(0.697)		-3.103	(1.133)	**
W1 physical activity	-0.207	(0.096)	*	-0.052	(0.148)	
W1 drinking	0.054	(0.068)		-0.124	(0.083)	
_constant	-10.229	(2.326)	***	-8.879	(2.879)	**

Table 17: -cont-

	White Female			White Male		
	Est	SE		Est	SE	
Wave 4 Age	0.086	(0.032)	**	0.109	(0.035)	**
Region: Midwest	0.256	(0.174)		0.598	(0.174)	***
Region: South	0.075	(0.172)		0.290	(0.175)	
Region: Northeast	1.053	(0.197)	***	0.791	(0.196)	***
HH: 2 parents	-0.649	(0.142)	***	-0.416	(0.152)	**
HH: single mother	-0.043	(0.187)		0.035	(0.188)	
HH: single father	0.335	(0.388)		-0.432	(0.323)	
HH: other	-1.100	(0.330)	***	-0.370	(0.425)	
Mother's education	0.058	(0.163)		0.018	(0.203)	
Mother's education^2	0.011	(0.014)		0.011	(0.017)	
Father's education	-0.076	(0.150)		-0.167	(0.161)	
Father's education^2	0.020	(0.013)		0.029	(0.014)	*
Log family income	-0.161	(0.446)		-0.636	(0.307)	*
Log family income^2	0.092	(0.066)		0.173	(0.048)	***
Lived environment	0.452	(0.132)	**	0.378	(0.142)	**
PVT score	0.027	(0.005)	***	0.031	(0.005)	***
GPA	1.424	(0.092)	***	1.213	(0.092)	***
Birth weight	0.005	(0.003)		0.004	(0.003)	
W1 SRH	0.240	(0.077)	**	0.163	(0.075)	*
W1 BMI	-0.024	(0.016)		-0.008	(0.015)	
W1 daily smoke	-0.196	(0.386)		-0.623	(0.482)	
W1 physical activity	0.044	(0.061)		0.126	(0.066)	
W1 drinking	0.024	(0.042)		-0.041	(0.039)	
_constant	-14.930	(1.400)	***	-14.630	(1.542)	***

Notes: Estimates are log odds from logit models using sample weights. The estimates are averages across 10 multiply imputed data sets with robust standard errors combined using Rubin's formula.

APPENDIX THREE: DESCRIPTIVE STATISTICS FOR COVARIATES (DEPRESSION)

Table 18: Descriptive Statistics for Covariates of Depression.

	Mean
Wave 4 Age	28.50
Region: West	0.14
Region: Midwest	0.31
Region: South	0.42
Region: Northeast	0.13
HH: 2 bio parents	0.51
HH: 2 parents	0.19
HH: single mother	0.22
HH: single father	0.03
HH: other	0.05
Mother's education	5.68
Father's education	5.58
Log family income	3.57
Lived environment	3.37
PVT score	102.18
GPA	2.80

Note: Based on 10 complete data sets. Weighted data.

APPENDIX FOUR: PROPENSITY SCORE MODELS (DEPRESSION)

Table 19: Propensity Score Models for Depression Analysis.

	Black Female			Black Male		
	Est	SE		Est	SE	
Wave 4 Age	-0.048	0.050		-0.021	0.066	
Region: Midwest	-0.321	0.344		0.336	0.446	
Region: South	0.036	0.292		0.170	0.389	
Region: Northeast	0.096	0.432		0.466	0.655	
HH: 2 parents	-0.734	0.272	**	-1.000	0.387	**
HH: single mother	-0.420	0.243		-0.560	0.329	
HH: single father	-0.646	0.533		0.637	0.910	
HH: other	-1.103	0.367	**	0.088	0.433	
Mother's education	0.131	0.230		-0.671	0.353	*
Mother's edu sq.	0.007	0.022		0.075	0.030	**
Father's education	-0.087	0.243		0.144	0.347	
Father's edu sq.	0.015	0.023		0.008	0.030	
Log family income	0.239	0.637		-0.120	0.639	
Log family inc sq.	0.015	0.105		0.021	0.113	
Lived environment	0.118	0.141		0.142	0.174	
PVT score	0.047	0.008	***	0.033	0.010	***
GPA	1.121	0.161	***	1.148	0.177	***
_cons	-9.182	2.073	***	-8.041	2.685	**

Table 19: -cont-

	White Female			White Male		
	Est	SE		Est	SE	
Wave 4 Age	0.094	0.030	**	0.095	0.032	**
Region: Midwest	0.259	0.175		0.610	0.176	***
Region: South	0.072	0.173		0.283	0.175	
Region: Northeast	1.079	0.196	***	0.846	0.194	***
HH: 2 parents	-0.647	0.141	***	-0.423	0.151	**
HH: single mother	-0.056	0.199		0.022	0.184	
HH: single father	0.190	0.378		-0.484	0.321	
HH: other	-1.201	0.330	***	-0.396	0.409	
Mother's education	0.057	0.161		0.067	0.207	
Mother's edu sq.	0.012	0.014		0.007	0.017	
Father's education	-0.054	0.174		-0.145	0.160	
Father's edu sq.	0.019	0.015		0.028	0.014	*
Log family income	-0.232	0.438		-0.641	0.285	*
Log family inc sq.	0.103	0.062		0.175	0.047	***
Lived environment	0.451	0.108	***	0.412	0.121	***
PVT score	0.027	0.005	***	0.028	0.005	***
GPA	1.455	0.090	***	1.266	0.090	***
_cons	-13.457	1.344	***	-12.834	1.443	***

Notes: Estimates are log odds from logit models using sample weights. The estimates are averages across 10 multiply imputed data sets with robust standard errors combined using Rubin's formula.

REFERENCES

Adair, Linda S., Barry M. Popkin, John S. Akin, David K. Guilkey, Socorro Gultiano, Judith Borja, Lorna Perez, Christopher W. Kuzawa, Thomas McDade, and Michelle J. Hindin. 2011. "Cohort Profile: The Cebu Longitudinal Health and Nutrition Survey." *International Journal of Epidemiology* 40(3):619-25.

Aldrich, John H., and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Beverly Hill, CA: Sage.

Allison, Paul D. 2002. *Missing Data*. Thousand Oaks, CA: Sage.

—. 2009. "Missing Data." Pp. 72-89 in *The SAGE Handbook of Quantitative Methods in Psychology*, edited by Roger E. Millsap and Alberto Maydeu-Olivares. Thousand Oaks, CA: Sage.

Almond, Douglas, and Janet Currie. 2011. "Killing Me Softly: The Fetal Origins Hypothesis." *Journal of Economic Perspectives* 25(3):153-72.

Arbuckle, James L. 1996. "Full Information Estimation in the Presence of Incomplete Data." Pp. 243-77 in *Advanced Structural Equation Modeling: Issues and Techniques*, edited by George A. Marcoulides and Randall E. Schumacker. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ashenfelter, Orley, and Alan B. Krueger. 1994. "Estimates of the Economic Return to Schooling from a New Sample of Twins." *American Economic Review* 84(5):1157-73.

Ashenfelter, Orley, and Cecilia Rouse. 1998. "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins." *Quarterly Journal of Economics* 113:253-84.

Asparouhov, Tihomir, and Bengt Muthén. 2010. "Chi-Square Statistics with Multiple Imputation." *Technical Report* http://www.statmodel.com.

Ballard, Jeanne L., Kathy Kazmaier Novak, and B. A. Marshall Driver. 1979. "A Simplified Score for Assessment of Fetal Maturation of Newly Born Infants." *Journal of Pediatrics* 95(5):769-74.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61:962-72.

Baraldi, Amanda N., and Craig K. Enders. 2010. "An Introduction to Modern Missing Data Analyses." *Journal of School Psychology* 48:5-37.

Bjelland, Ingvar, Steinar Krokstad, Arnstein Mykletun, Alv A. Dahl, Grethe S. Tell, and K. Tambs. 2008. "Does a Higher Educational Level Protect Against Anxiety and Depression? The HUNT Study." *Social Science & Medicine* 66:1334-45.

Blau, Peter M., and Otis Dudley Duncan. 1967. *The American Occupational Structure*. New York: Wiley.

Boardman, Jason D., Jarron M. Saint Onge, Richard G. Rogers, and Justin T. Denney. 2005. "Race Differentials in Obesity: The Impact of Place." *Journal of Health and Social Behavior* 46(3):229-43.

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

Bollen, Kenneth A., Linda S. Adair, and Mark D. Noble. 2011. "Measurement Error and Missing Data in Gestational Age, Birth Weight, and Birth Length in Filipino Infants Analyzed with Structural Equation Models." *manuscript*.

Bollen, Kenneth A., and Shawn Bauldry. 2011. "Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates." *Psychological Methods* 16(3):264-84.

Brand, Jennie E. 2010. "Civic Returns to Higher Education: A Note on Heterogeneous Effects." *Social Forces* 89(2):417-34.

Brand, Jennie E., and Dwight Davis. 2011. "The Impact of College Education on Fertility: Evidence for Heterogeneous Effects." *Demography*.

Brand, Jennie E., and Yu Xie. 2010. "Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Review* 75:272-302.

Brookhart, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. 2006. "Variable Selection for Propensity Score Models." *American Journal of Epidemiology* 163(12):1149-56.

Browne, Michael W., and Robert Cudeck. 1993. "Alternative Ways of Assessing Model Fit." Pp. 136-62 in *Testing Structural Equation Models*, edited by Kenneth A. Bollen and J. Scott Long. Newbury Park, CA: Sage.

Card, David. 1999. "The Causal Effect of Education on Earnings." Pp. 1801-63 in *Handbook of Labor Economics*, edited by Orley Ashenfelter and David Card. Amsterdam: Elsevier.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. Pacific Grove, CA: Duxbury.

Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network Over 32 Years." *New England Journal of Medicine* 357:370-79.

—. 2008. "The Collective Dynamics of Smoking in a Large Social Network." *New England Journal of Medicine* 358:2249-58.

Cockerham, William C. 1990. "A Test of the Relationship Between Race, Socioeconomic Status, and Psychological Distress." *Social Science & Medicine* 31(12):1321-26.

Collins, Linda M., Joseph L. Schafer, and Chi-Ming Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6(4):330-51.

Cutler, David M., and Edward L. Glaeser. 2007. "Social Interactions and Smoking." in *Harvard Institute of Economic Research Discussion Paper Number 2153*.

Cutler, David M., and Adriana Lleras-Muney. 2008. "Education and Health: Evaluating Theories and Evidence." Pp. 29-60 in *Making Americans Healthier: Social and Economic Policy as Healthy Policy*, edited by Robert F. Schoeni, James S. House, George A. Kaplan, and Harold Pollack. New York: Russell Sage Foundation.

—. 2010. "Understanding Differences in Health Behaviors by Education." *Journal of Health Economics* 29:1-28.

Dahly, Darren L., Linda S. Adair, and Kenneth A. Bollen. 2009. "A Structural Equation Model of the Developmental Origins of Blood Pressure." *International Journal of Epidemiology* 38:538-48.

Demirtas, Hakan. 2004. "Simulation Driven Inferences for Multiply Imputed Longitudinal Datasets." *Statistica Neerlandica* 58(4):466-82.

Dunn, L., and L. Dunn. 1981. *PPVT-R Manual*. Circle Pines, MN: American Guidance Service.

Dupre, Matthew E. 2007. "Educational Differences in Age-Related Patterns of Disease: Reconsidering the Cumulative Disadvantage and Age-As-Leveled Hypotheses." *Journal of Health and Social Behavior* 48:1-15.

Elo, Irma T. 2009. "Social Class Differentials in Health and Mortality: Patterns and Explanations in Comparative Perspective." *Annual Review of Sociology* 35:553-72.

Enders, Craig K. 2008. "A Note on the Use of Missing Auxiliary Variables in Full Information Maximum Likelihood-Based Structural Equation Models." *Structural Equation Modeling* 15:434-48.

—. 2010. *Applied Missing Data Analysis*. New York: The Guilford Press.

Farmer, Melissa M., and Kenneth F. Ferraro. 2005. "Are Racial Disparities in Health Conditional on Socioeconomic Status?" *Social Science & Medicine* 60(1):191-204.

Ferraro, Kenneth F., and Melissa M. Farmer. 1999. "Utility of Health Data from Social Surveys: Is there a Gold Standard for Measuring Morbidity?" *American Sociological Review* 64(2):303-15.

Fields, Larry E., Vicki L. Burt, Jeffrey A. Cutler, Jeffrey Hughes, Edward J. Roccella, and Paul Sorlie. 2004. "The Burden of Adult Hypertension in the United States 1999 to 2000: A Rising Tide." *Hypertension* 44:398-404.

Fischer, Claude S., and Michael Hout. 2006. *Century of Difference: How America Changed in the Last One Hundred Years*. New York: Russell Sage Foundation.

Freese, Jeremy, and Karen Lutfey. 2011. "Fundamental Causality: Challenges of an Animating Concept for Medical Sociology." Pp. 67-81 in *Handbook of the Sociology of health, Illness, and Healing: A Blueprint for the 21st Century*, edited by Bernice A. Pescosolido, Jack K. Martin, Jane D. McLeod, and Anne Rogers. New York: Springer.

Goesling, Brian. 2007. "The Rising Significance of Education for Health?" *Social Forces* 85(4):1621-43.

Gottfredson, Linda S. 2004. "Intelligence: Is It the Epidemiologists' Elusive "Fundamental Cause" of Social Class Inequalities in Health?" *Journal Of Personality and Social Psychology* 86:174-99.

Graham, John W. 2003. "Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models." *Structural Equation Modeling* 10:80-100.

Griliches, Zvi. 1977. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45:1-22.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5):646-75.

Guo, Shenyang, and Mark W. Fraser. 2010. *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: Sage.

Harris, Kathleen Mullan. 1999. "The Health Status and Risk Behavior of Adolescents in Immigrant Families." Pp. 286-347 in *Children of Immigrants: Health, Adjustment, and Public Assistance*, edited by Donald J. Hernandez. Washington, D.C.: National Academy Press.

Harris, Kathleen Mullan, Carolyn T. Halpern, Eric A. Whitsel, Jon M. Hussey, Joyce W. Tabor, Pamela P. Entzel, and J. Richard Udry. 2009. "The National Longitudinal Study of Adolescent Health: Research Design." URL: http://www.cpc.unc.edu/projects/addhealth/design.

Hauser, Robert M., and Thomas N. Daymount. 1977. "Schooling, Ability, and Earnings: Cross-Sectional Findings 8 to 14 Years after High School Graduation." *Sociology of Education* 50:182-206.

Heckman, James J. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35:1-97.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88:389-432.

Hirschi, Travis, and Michael R. Gottfredson (Eds.). 1994. *The Generality of Deviance*. New Brunswick, NJ: Transaction Publishers.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945-70.

House, James S., Karl R. Landis, and Debra Umberson. 1988. "Social Relationships and Health." *Science* 241:540-45.

House, James S., Paula M. Lantz, and Pamela Herd. 2005. "Continuity and Change in the Social Stratification of Aging and Health Over the Life Course: Evidence From a Nationally Representative Longitudinal Study From 1986 to 2001/2002 (Americans' Changing Lives Study)." *Journals of Gerontology (Social Sciences)* 60B(Special Issue II):15-26.

Hout, Michael. 1988. "More Universalism, Less Structural Mobility: The American Occupational Structure in the 1980s." *American Journal of Sociology* 93:1358-400.

Hout, Michael, and Thomas A. DiPrete. 2006. "What Have We Learned: RC28's Contributions to Knowledge about Social Stratification." *Research in Social Stratification and Mobility* 24:1-20.

Hudson, Darrell L., H. W. Neighbors, A. T. Geronimus, and J. S. Jackson. 2012. "The Relationship Between Socioeconomic Position and Depression Among a US Nationally Representative Sample of African Americans." *Social Psychiatry and Psychiatric Epidemiology* 47(3):373-81.

Ibrahim, Joseph G., Stuart R. Lipsitz, and Nick Horton. 2001. "Using Auxiliary Data for Parameter Estiamtion with Non-Ignorably Missing Outcomes." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50(3):361-73.

Idler, Ellen L., and Yael Benyamini. 1997. "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." *Journal of Health and Social Behavior* 38:21-37.

Jackson, Pamela Braboy, and Jason Cummings. 2011. "Health Disparities and the Black Middle Class: Overview, Empirical Findings, and Research Agenda." Pp. 383-410 in *Handbook of the Sociology of Health, Illness, and Healing: A Blueprint for the 21st Century*, edited by Bernice A. Pescosolido, Jack K. Martin, Jane D. McLeod, and Anne Rogers. New York: Springer.

Jylhä, Marja. 2009. "What Is Self-Rated Health and Why Does It Predict Mortality? Towards a Unified Conceptual Model." *Social Science & Medicine* 69(3):307-16.

Jylhä, Marja, Stefano Volpato, and Jack M. Guralnik. 2006. "Self-Rated Health Showed a Graded Association with Frequently Used Biomarkers in a Large Population Sample." *Journal of Clinical Epidemiology* 59:465-71.

Kalton, Graham, and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19(2):81-97.

Kawachi, Ichiro, S. V. Subramanian, and Daniel Kim (Eds.). 2008. *Social Capital and Health*. New York: Springer.

Kessler, Ronald C., Katherine A. McGonagle, Shanyang Zhao, Christopher B. Nelson, Michael Hughes, Suzann Eshleman, Han-Ulrich Wittchen, and Kenneth S. Kendler. 1994. "Lifetime and 12-Month Prevalence of DMS-III-R Psychiatric Disorders in the United States." *Archives of General Psychiatry* 51:8-19.

Keyes, Corey L. M., and Sherryl H. Goodman (Eds.). 2006. *Women and Depression: A Handbook for the Social, Behavioral, and Biomedical Sciences*. New York: Cambridge University Press.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49-69.

Kitagawa, Evelyn M., and Philip M. Hauser. 1973. *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge: Harvard University Press.

Kreuter, Frauke, and Kristen Olson. 2011. "Multiple Auxiliary Variables in Nonresponse Adjustment." *Sociological Methods & Research* 40(2):311-32.

Lantz, Paula M., James S. House, Richard P. Mero, and David R. Williams. 2005. "Stress, Life Events, and Socioeconomic Disparities in Health: Results from the Americans' Changing Lives Study." *Journal of Health and Social Behavior* 46:274-88.

Lawes, Carlene M. M., Stephen Vander Hoorn, and Anthony Rodgers. 2008. "Global Burden of Blood-Pressure-Related Disease, 2001." *Lancet* 371(9623):1513-18.

Link, Bruce G., and Jo Phelan. 1995. "Social Conditions as Fundamental Causes of Disease." *Journal of Health and Social Behavior* 35:80-94.

Link, Bruce G., Jo C. Phelan, Richard Miech, and Emily Leckman Westin. 2008. "The Resources That Matter: Fundamental Social Causes of Health Disparities and the Challenge of Intelligence." *Journal of Health and Social Behavior* 49:72-91.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken, NJ: Wiley.

Little, Roderick J., and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31(2):161-68.

Liu, Hui, and Robert A. Hummer. 2008. "Are Educational Differences in U.S. Self-Rated Health Increasing?: An Examination by Gender and Race." *Social Science & Medicine* 67:1898-906.

Lorant, V., D. Deliège, W. Eaton, A. Robert, P. Philippot, and M. Ansseau. 2003. "Socioeconomic Inequalities in Depression: A Meta-Analysis." *American Journal of Epidemiology* 157(2):98-112.

Luo, Ye, and Linda J. Waite. 2005. "The Impact of Childhood and Adult SES on Physical, Mental, and Cognitive Well-Being in Later Life." *Journals of Gerontology (Social Sciences)* 60B(2):S93-S101.

Lynch, Scott M. 2003. "Cohort and Life-Course Patterns in the Relationship Between Education and Health: A Hierarchical Approach." *Demography* 40(2):309-31.

—. 2006. "Explaining Life Course and Cohort Variation in the Relationship between Education and Health: The Role of Income." *Journal of Health and Social Behavior* 47:324-38.

Meng, Xiao-Li, and Donald B. Rubin. 1992. "Performing Likelihood Ratio Tests with Multiply-Imputed Data Sets." *Biometrika* 79:103-11.

Miech, R. A., and M. J. Shanahan. 2000. "Socioeconomic Status and Depression Over the Life Course." *Journal of Health and Social Behavior* 41(2):162-76.

Miech, Richard A., Avshalom Caspi, Terrie E. Moffitt, Bradley R. Entner Wright, and Phil A. Silva. 1999. "Low Socoieconomic Status and Mental Disorders: A Longitudinal Study of Selection and Causation during Young Adulthood." *American Journal of Sociology* 104(4):1096-131.

Mirowsky, John, and Catherine E. Ross. 1998. "Education, Personal Control, Lifestyle and Health: A Human Capital Hypothesis." *Research on Aging* 20(4):415-49.

—. 2003. *Education, Social Status, and Health*. New York: Aldine De Gruyter.

—. 2008. "Education and Self-Rated Health: Cumulative Advantage and Its Rising Importance." *Research on Aging* 30(1):93-122.

Mokdad, Ali H., James S. Marks, Donna F. Stroup, and Julie L. Gerberding. 2004. "Actual Causes of Death in the United States, 2000." *Journal of the American Medical Association* 291(10):1238-45.

Morgan, Stephen L., and David J. Harding. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls of Theory and Practice." *Sociological Methods & Research* 35(1):3-60.

Morgan, Stephen L., and Jennifer J. Todd. 2008. "A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects." *Sociological Methodology* 38:231-81.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Murphy, Jane M. 2011. "Symptom Scales and Diagnostic Schedules in Adult Psychiatry." Pp. 199-219 in *Textbook in Psychiatric Epidemiology, 3rd ed.*, edited by Ming T. Tsuang, Mauricio Tohen, and Peter Jones. New York: Wiley.

Mustillo, Sarah A. forthcoming. "The Effects of Auxiliary Variables on Coefficient Bias and Efficiency in Multiple Imputation." *Sociological Methods & Research*.

Muthén, Linda K., and Bengt Muthén. 1998-2010. *Mplus User's Guide*. Los Angeles: Muthén & Muthén.

Nguyen, Quynh C., Joyce W. Tabor, Pamela P. Entzel, Yan Lau, Chirayath M. Suchindran, Jon M. Hussey, Carolyn T. Halpern, Kathleen Mullan Harris, and Eric A. Whitsel. 2011. "Discordance in National Estimates of Hypertension Among Yound Adults." *Epidemiology* 22(4):532-41.

Pampel, Fred C., Patrick M. Krueger, and Justin T. Denney. 2010. "Socioeconomic Disparities in Health Behaviors." *Annual Review of Sociology* 36(1):349-70.

Perreira, Krista M., Natalia Deeb-Sossa, Kathleen Mullan Harris, and Kenneth A. Bollen. 2005. "What Are We Measuring? An Evaluation of the CES-D Across Race/Ethnicity and Immigrant Generation." *Social Forces* 83(4):1567-601.

Preston, Samuel H., and Irma T. Elo. 1995. "Are Educational Differentials in Adult Mortality Increasing in the United States?" *Journal of Aging and Health* 7:476-96.

Radloff, Lenore S. 1977. "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population." *Applied Psychological Measurement* 1(3):385-401.

Raftery, Adrian. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-63.

Riolo, Stephanie, Tuan Anh Nguyen, John F. Greden, and Cheryl A. King. 2005. "Prevalence of Depression by Race/Ethnicity: Findings from the National Health and Nutrition Examination Survey III." *American Journal of Public Health* 95(6):998-1000.

Robins, James M., and Andrea Rotnitzky. 1997. "Comment on 'Inference for Semiparametric Models: Some Questions and an Answer' by Bickel and Kwon." *Statistica Sinica* 11:920-36.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41-55.

Ross, Catherine E., and John Mirowsky. 1989. "Explaining the Social Patterns of Depression: Control and Problem Solving - Or Support and Talking." *Journal of Health and Social Behavior* 30:206-19.

—. 2006. "Sex Differences in the Effect of Education on Depression: Resource Multiplication or Resource Substitution?" *Social Science & Medicine* 63:1400-13.

—. 2010. "Why Education Is the Key to Socioeconomic Differentials in Health." Pp. 33-51 in *Handbook of Medical Sociology (6th ed)*, edited by Chloe E. Bird, Peter Conrad, Allen M. Fremont, and Stefan Timmermans. Nashville, TN: Vanderbilt University Press.

Rubin, Donald B. 1973. "Matching to Remove Bias in Observational Studies." *Biometrics* 29:159-83.

—. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688-701.

—. 1976. "Inference and Missing Data." *Biometrika* 63:581-92.

—. 1987. *Multiple Imputation for Survey Nonresponse*. New York: Wiley.

—. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91:473-89.

—. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127(8S):757-63.

—. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Savalei, Victoria, and Peter M. Bentler. 2009. "A Two-Stage Approach to Missing Data: Theory and Application to Auxiliary Variables." *Structural Equation Modeling* 16:477-97.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Missing Data*. New York: Chapman and Hall.

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7(2):147-77.

Schafer, Joseph L., and Maren K. Olsen. 1998. "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective." *Multivariate Behavioral Research* 33(4):545-71.

Schwartz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6(2):461-64.

Seeman, Teresa E. 1996. "Social Ties and Health: The Benefits of Social Integration." *Annals of Epidemiology* 6(5):442-51.

Shuey, Kim M., and Andrea E. Willson. 2008. "Cumulative Disadvantage and Black-White Disparities in Life-Course Health Trajectories." *Research on Aging* 30(2):200-25.

Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27:325-53.

Smith, James P. 2007. "The Impact of Socioeconomic Status on Health over the Life-Course." *Journal of Human Resources* 42(4):739-64.

Smith, Kristen P., and Nicholas A. Christakis. 2008. "Social Networks and Health." *Annual Review of Sociology* 34:405-29.

Snyder, Thomas D., and Sally A. Dillow. 2011. "Digest of Education Statistics 2010." Washington, D.C.: National Center for Education Statistics.

Somervell, Philip D., Philip J. Leaf, Myrna M. Weissman, Dan G. Blazer, and Martha Livingston Bruce. 1989. "The Prevalence of Major Depression in Black and White Adults in Five United States Communities." *American Journal of Epidemiology* 130(4):725-35.

Splawa-Neyman, Jerzy. [1923] 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (translated, with discussion)." *Statistical Science* 5(4):465-80.

StataCorp. 2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp, LP.

Steiger, James H., and John C. Lind. 1980. "Statistically-Based Tests for the Number of Common Factors." in *Annual Meeting of the Psychometric Society*. Iowa City, IA.

Strine, Tara W., Ali H. Mokdad, Lina S. Balluz, Olinda Gonzalez, Raquel Crider, Joyce T. Berry, and Kurt Kroenke. 2008. "Depression and Anxiety in the United States: Findings from the 2006 Behavioral Risk Factor Surveillance System." *Psychiatric Services* 59(12):1383-90.

Stuart, Alan, Keith Ord, Steven Arnold, Anthony O'Hagan, and Jon Forster. 2005. *Kendall's Advanced Theory of Statistics*. New York: Oxford University Press.

Turner, Jay R., Blair Wheaton, and Donald A. Lloyd. 1995. "The Epidemiology of Social Stress." *American Sociological Review* 60:104-25.

Turner, R. Jay, and Donald A. Lloyd. 1999. "The Stress Process and the Social Distribution of Depression." *Journal of Health and Social Behavior* 40(4):374-404.

von Hippel, Paul T. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology* 37:83-117.

Wheaton, Blair. 1978. "The Sociogenesis of Psychological Disorder: Reexamining the Causal Issues with Longitudinal Data." *American Sociological Review* 43(3):383-403.

—. 1980. "The Sociogenesis of Psychological Disorder: An Attributional Theory." *Journal of Health and Social Behavior* 21(2):100-24.

Williams, David R., and Chiquita Collins. 1995. "US Socioeconomic and Racial Differences in Health: Patterns and Explanations." *Annual Review of Sociology* 21:349-86.

Williams, David R., Hector M. Gonzalez, Harold Neighbors, Randolph Nesse, Jamie M. Abelson, Julie Sweetman, and James S. Jackson. 2007. "Prevalence and Distribution of Major Depressive Disorder in African Americans, Caribbean Blacks, and Non-Hispanic Whites: Results from the National Survey of American Life." *Archives of General Psychiatry* 64:305-15.

Willson, Andrea E., Kim M. Shuey, and Glen H. Elder. 2007. "Cumulative Advantage Processes as Mechanisms of Inequality in Life Course Health." *American Journal of Sociology* 112(6):1886-924.

World Health Organization. 2004. *The World Health Report 2004: Changing History*. Geneva: WHO.

Yoo, Jin Eun. 2009. "The Effect of Auxiliary Variables and Multiple Imputation on Parameter Estimation in Confirmatory Factor Analysis." *Educational and Psychological Measurement* 69(6):929-47.

Yuan, Ke-Hai, and Peter M. Bentler. 2000. "Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data." *Sociological Methodology* 30:165-200.

Zheng, Hui, and Kenneth C. Land. 2012. "Composition and Decomposite in US Gender-Specific Self-Reported Health Disparities, 1984-2007." *Social Science Research* 41:477-88.