Center for Demography and Ecology

University of Wisconsin-Madison

# Sample Selection Bias in the Pathways to Adult Health Inequalities

Robert G. White

Alberto Palloni

CDE Working Paper No. 2009-11

# Sample Selection Bias in the Pathways to Adult Health Inequalities

May 5, 2009

Robert G. White[†]
Center for Demography and Ecology
Department of Sociology
University of Wisconsin, Madison

Alberto Palloni
Center for Demography and Ecology
Department of Sociology
University of Wisconsin, Madison

## Abstract

Sample selection bias is a chronic problem in longitudinal studies that is particularly problematic for studies concerning the relationship between health and socio-economic status. This paper adopts two alternate methods for handling sample selection bias attributable to survey attrition and item non-response. Both methods are applied to examine the magnitude of bias in the effects of childhood cognition and behavior on the adult socio-economic gradient in health. A method for sample selection correction with multiple imputation for item non-response is implemented to account for different sources of sample selection bias over time. Estimates of a life course model of health and socioeconomic attainment demonstrate that sample selection bias inflates estimates of socioeconomic gradients. The proposed correction for sample selection bias also suggests that the effects of early child non-cognitive skills rather than cognitive skills may play an important role in the early life origins of adult socioeconomic gradients.

[†] Contact address: 1180 Observatory Drive, Center for Demography and Ecology University of Wisconsin, Madison, WI 53706. Email: rwhite@ssc.wisc.edu.

## I. Introduction

Socio-economic gradients in adult health outcomes have been among the most regularly observed empirical phenomena in the social sciences. From the 19th century (Chadwick 1842) through early studies of mortality and education (Kitagawa and Hauser 1973) to the growth in social epidemiology following the Whitehall II study (Marmot et al. 1991), alternate measures of good health over different periods of the life course consistently correspond with higher socio-economic status. Many explanations of these relationships update Kitigawa and Hauser's (1973) initial explanation of health gradients as reflecting differences in resources which accrue with educational attainment and that influence contemporaneous health. Such resources may include individual behaviors, such as smoking, drinking and body mass index (Kaplan et al. 1987; Preston and Taubman 1994; Ecob and Davey Smith 1999; Currie and Moretti 2003) in addition to the effects of lower socioeconomic status on both health services utilization (Link and Phelan 1995; Ross and Wu 1995, Gornick et al. 1996; Kirby and Kaneda 2006) and the duration of exposure to stress (Link and Phelan 1995; Stafford and Marmot 2003; Dowd and Goldman 2006). An alternate set of explanations reverse the direction of causality and attribute adverse health to changes in wealth and labor market outcomes which lower trajectories of socioeconomic attainment (Shorrocks 1975). Given the importance of the mid- and late-stages of lifetime labor market participation for career advancement and asset accumulation, health shocks in middle adulthood may pose particularly large risks for future socioeconomic status (Adams et al. 2003; Smith 1999, 2004).

While there is wide agreement on the importance of both sets of causal pathways, it is also increasingly evident that the origins of socioeconomic gradients in health precede many of these contemporaneous and near-term relationships. Growing attention across the medical and social sciences to earlier periods in the life course have extended the time period for considering cumulative effects of low socioeconomic status and risky health behaviors to childhood. Hypotheses tracing adult chronic conditions to prenatal and infant health (Barker et al. 1993; Barker 1995) have given rise to life course approaches in epidemiology (Kuh et al. 2003) that increasingly link birthweight and infant growth to adult risks for diabetes and cardiovascular disease. Evidence from twins studies (Conley et al. 2003; Behrman and Rosenzweig 2004; Black et al. 2007; Oreopoulos et al. 2008) and quasi-experimental methods (Almond 2006; Bleakley 2007; Shillingford 2008; Clarke et al. 2008) also point to early health for similarly long lasting effects in schooling outcomes and socioeconomic attainment. The accumulating evidence from these approaches suggest dynamic ties between health and socioeconomic attainment over a wide span of the life course.

An important challenge for any study that addresses such a large number of hypothesized pathways over different periods of the life course is minimizing the loss of sample representativeness due to survey attrition and item non-response. These concerns are especially acute in longitudinal studies. Both item non-response and attrition have been associated with socioeconomic background, education, occupation, income (Grovers and Couper 1988; Bound and Krueger 1991) and other characteristics that are among the main pathways among the hypothesized links between socio-economic attainment and health. Refreshment sampling in later survey rounds that oversample individuals along the predictors of subsequent non-response may only complicate the nature of this bias. In particular, without knowledge of the duration in low socio-economic status, changes in the effects of socio-economic status over time will result in further bias from over sampling low socio-economic status individuals. The directions of these biases and their changes over time are not easily predicted. Moreover, correcting such bias is further complicated by changing probabilities of survey non-response over time that may also depend on socio-economic status.

This paper implements a procedure for simultaneously addressing item missingness and attrition.  A conventional Markov Chain Monte Carlo (MCMC) procedure for multiple imputation is implemented in combination with a weighting method for sample selection correction.  This procedure aims to relax the demands for multiple imputation by conditioning imputation on a weighting scheme that corrects selection bias related to attrition.  The procedure is implemented in a life course model of early health and adult economic attainment using the National Child Development Study (NCDS) 1958 birth cohort from the United Kingdom.  This dataset has a rich collection of measures for studying development, health and socioeconomic attainment over the life course and spans the life course from birth to age 46.

The paper proceeds in section two by developing a method for sample selection correction that combines a weighting scheme and a MCMC method for multiple imputation. Section three proposes a model for linking early health to adult socioeconomic gradients in health that permits examining the consequences of the proposed method for life course studies of health and socioeconomic attainment.  Section four describes the NCDS and illustrates the select process of survey participation that threatens unbiased estimates of life course models. Section five implements the proposed procedure and presents results.

## II. Correcting Sample Selection Bias in Longitudinal Data

There are a range of methods for mitigating the effects of incomplete data.  Data that is incomplete due to item missingness may be addressed with methods for data imputation that emphasize alternate assumptions about the processes influencing missingness in order to impute missing values.  Incomplete data that is attributable to survey attrition is often managed with sample selection correction models which are implemented by variants of Heckman's (1979) selection model or weighting schemes that correct for survivor bias.  The presence of both item non-response and attrition implies a disparate set of causes underlying the observed pattern of missingness. For instance, behavioral responses to participating in a survey which relate to the opportunity costs of time and geographic mobility characterize an attrition process that may share little in common with the factors underlying item missingness in measures of individual income. However, the information that is used in the separate methods for correcting these two classes of incomplete data often makes use of separate sets of information. Many methods for studying attrition are limited to using information from sampling frames or the data collection process while imputation methods include the additional information collected in the survey. Depending on the distribution of item missingness and attrition, the presence of multiple causes of missing data presents an opportunity to take advantage of the information revealed in these different processes to correct the resulting threats to unbiased estimation.

Despite the acute concerns over the scope of both item missingness and attrition in longitudinal studies, methods for addressing incomplete data in longitudinal studies often adopt partial solutions. While methods for multiple imputation have been recently introduced in epidemiologic studies with longitudinal data, these methods are commonly implemented to manage both types of incomplete data without accounting for the different underlying reasons for incomplete data. By contrast, traditional weighting methods are often implemented in the absence of procedures for correcting item missingness. Yet, the benefits of correcting selection bias that is related to attrition may be severely limited without mitigating the effects of missing data. Adapting both of these ex-post methods for managing incomplete data in order to simultaneously account for the biases due to item missingness and attrition presents a feasible approach for improving efficiency and minimizing the consequences of selection bias.

The proposed method incorporates a weighting scheme for correcting sample selection bias due to attrition with a method for multiple imputation. The first step in this procedure is defining a weighting scheme for attrition. A multiple imputation procedure is then implemented with the resulting weighted data in order to condition the imputation of missing values on the selection correction.

*Sample Selection Correction*

An inverse probability weighted (IPW) estimator is adopted and applied to the ordered probit model (Woolridge 2007) for health at age 46 that is developed in the next section. This estimator is implemented by first defining a sample selection model for estimating predicted probabilities of survey participation.[1] These probabilities are defined for each wave of the survey. A general formulation for describing participation in a survey over time relates individual characteristics at the baseline survey and during the course of the study to individual participation:

$$f(p_{it}) = \sum_{c=1}^{C} \beta_c x_{ci} + \sum_{m=1}^{M} \sum_{k=0}^{t-1} \lambda_{m,t-k} z_{mi,t-k} \tag{1}$$

The probability of participating ($p_{it}$) is defined for baseline individuals $i=1,\ldots, n_0$ at any wave $t$ in the survey. This probability is equivalent to the average response rate, $E(r_{it})$, where $r_{it}$ = 1 if individuals participate in the survey and 0 otherwise. Two sets of covariates distinguish the selection process. Individual characteristics $x_{ci}$ measure attributes at baseline which reflect a stable underlying propensity to participate. Individual characteristics $z_{mi,t-k}$ for m=1,…,M are measured from wave $t$ to baseline and capture changes over time in both the types of measures that are available and in individual characteristics. A central assumption in this framework is that the selection process for participation in the survey is a function of observed measures.[2]

[1] The IPW estimator is described in Appendix 1.

[2] See the appendix for further discussion of this assumption.

Longitudinal data that spans distinct life course transitions are ill-suited for the selection model in Equation 1. One main problem is that respondents are subject to change over long or critical periods of the life course that may be consequential for response probabilities. The above formulation is often estimated with a random effects model (Woolridge 2007). However, longitudinal studies that begin in childhood, such as the NCDS, are particularly vulnerable to the consequences of child development for obtaining stable measures of individual characteristics. Changes in individual characteristics and personality are likely to have time varying effects on the probability of participation that are correlated with the effects of the standard measures of individual development and socioeconomic attainment. In the absence of a reliable set of measures in $x_{ci}$, Equation 1 reduces to the set of time varying characteristics that is less easily adapted for a random effects estimation. The second related problem is that there are few variables in longitudinal surveys with constant measures over the life course. Health measures such as weight, height and self-reported health assessments have maintained reasonable constant measures over most longitudinal studies. By contrast, measures of schooling achievement, behavior and indicators of socioeconomic status often undergo extensive changes over the course of a longitudinal study.

An alternative approach estimates a variant of Equation 1 for each wave of the survey. In this case, the sample at risk of attrition is the sample in wave *t-1,* rather than the baseline sample.

$$f(p_{it}) = \beta_0 + \sum_{m=1}^{M}\sum_{k=0}^{t-1}\lambda_{m,t-k}z_{mi,t-k} + \mathrm{I}\{t > 2\} * \sum_{k=2}^{t-1}\alpha_{t-k}r_{i,t-k} \qquad (2)$$

Equation 2 includes current and prior measures of individual characteristics in $z_{mi,t-k}$ and allows the variables measuring these changes to vary over time. Indicators for participation in prior waves are also included as controls to account for the potentially distinct probabilities for participation between respondents who repeatedly drop-out and return to the sample and individuals for drop-out and never return.

Equation 2 is estimated as a set of probits for participation across the seven waves of the NCDS. The second step in implementing the IPW estimator is applying the resulting predicted probabilities to define probability weights that may be incorporated in the maximum likelihood estimation of the ordered probit model for health. The inverse of the predicted probability of participation estimated from Equation 2 is introduced to the estimated model as a set of probability weights for the full dataset. Since estimation of a life course model includes measures from across the waves of the survey, the weights are cumulatively constructed across the participation probits for each wave (Contoyannis, Jones and Rice 2004).

### *Multiple Imputation*

The multiple imputation procedure adapts a Markov Chain Monte Carlo method for imputation (Rubin 1976, 1987) with a sequence of regressions for multiply imputing missing values (Raghunathan et al. 2001). The central assumption in this procedure is that missingness is ignorable after accounting for attrition. This requires that missingness for each variable does not depend on the value of the missing item after conditioning on the observed values of other available covariates in the data. Any parameters which may govern missingness may also not be related to the parameters in the models of interest. Irrespective of whether the pattern of missingness in the data is monotonic, each variable with item missingness is estimated using a large set of variables. This set includes the full set of covariates from the estimated model of interest. An additional auxiliary set of variables is included that is determined by a sequence of multivariate analyses for item non-response probabilities in the main covariates of interest. For each covariate with missingness, a sweep over a large set of variables available in the NCDS was undertaken in a model search exercise that aided in identifying a corresponding set of auxiliary variables. Each variable in the resulting dataset is then estimated with an appropriate model depending on whether the variable is categorical, count or continuous. Normalizing transformations are undertaken for several variables where necessary. The estimated coefficients are used to generate predicted values which are then used in the estimations for the remaining set of variables. Assuming that the chosen covariates may be characterized by a multivariate normal distribution, stochastic variation may be introduced into the predicted values which corresponds to the predicted error distribution for each regression. This process is continued until all variables with missingness have predicted values.

The uncertainty in model parameters is incorporated by repeating these predictions multiple times and making draws from the posterior distribution of the parameters for all the underlying models. For each set of model estimates, the result is an iterative process between random draws of parameters (conditional on the data) and random draws of the missing data (conditional on parameters). This iterative process continues until the values of parameters stabilize, reflecting convergence in the estimated set of model parameters. This entire procedure is repeated one thousand times to generate a single dataset. The optimal number of datasets created by this procedure is determined by the convergence of the estimated between-sample variance of the main coefficients of interest.

All of the empirical analysis is undertaken with a resulting set of ten datasets. Consequently, descriptive statistics reflect sample statistics that are averaged across these multiple datasets. Variances for all statistics are corrected to account for both within-sample and between-sample variances. Estimating models with these multiple datasets requires similar corrections for estimated coefficients and variances to account for the within- and between-sample variances. While multiply imputed values are generated for missingness in all variables, the imputed values for health and socioeconomic status at age 46 are not included in the sample used for the analysis.

## III. Health and Socioeconomic Attainment over the Life Course

This method for mitigating the effects of attrition and item non-response is implemented with an empirical framework that relates early health with the inputs for lifetime socioeconomic attainment. A model of health attainment over the life course is developed to describe early health effects in socioeconomic and health attainment. The proposed model extends Grossman's (1971) original formulation of health production by adapting Cunha and Heckman's (2008) framework for human capabilities development and emphasizing cognitive and non-cognitive skills in the accumulation of health and socioeconomic attainment. The model builds upon Todd and Wolpin's (2007) general formulation of cumulative attainment and remains agnostic about the precise processes underlying skills development.[3] Individual health status at age 46 depends on trajectories of health and socioeconomic attainment over the life course.

$$H_x = H_x(h_o, S_{x-1}, I_{x-1}) \qquad (3)$$

$S_x$ and $H_x$ represent individual socioeconomic position and health status at age x. $I_{x-1}$ is a vector of human capital and health status inputs in the age interval (x-1, x), and $h_o$ is a vector of traits that individuals acquire at birth which are beneficial for maintaining good health, but which are unobserved. Inputs may assume negative values in the case of adverse health shocks like negative exposures and trauma. Such a general formulation may include unobserved inputs and traits such as parents' time investments and genetic frailty and is readily adaptable to different

[3] The full theoretical framework is developed in Palloni, White and Milesi (2009).

datasets varying in the availability of measures of inputs, attainment, traits and proxy indicators of traits as well as in the frequencies of observations over the life cycle.

Adult health in the NCDS is measured with respondent's self assessments of overall health condition. This measure is reported by a four point scale reporting excellent, good, fair and poor health. This four point measure reflects a latent measure of underlying health and is estimated with an ordered probit approximation of Equation 3:

$$P_{ij} = P(H_i = j) = \Phi(\mu_j - X_i\beta - \varepsilon_i) - \Phi(\mu_{j-1} - X_i\beta - \varepsilon_i) \qquad (4)$$

The main measures of attainment are social class at age 42 and self-reported health status at age 46. The lagged effect of socioeconomic status is adopted in order to mitigate the possible reverse effects of health on socioeconomic status. Given the increasing risk for adverse health events with age, limiting the observations of adult attainment to measures in middle adulthood implies smaller estimates of adult health gradients during this period of adulthood. However, truncating the life course at age 46 also further minimizes the potential bias due to the likely increase in effects of health on socio-economic status and heterogeneity during late adulthood. Given the changing rate of women's labor force participation in the United Kingdom during the 1980s, the sample is limited to men in order to minimize the consequences of corresponding changes in the measurement of individual socioeconomic status during adulthood.[4]

An indicator of low birthweight is adopted as the first measure for early child health. Birth weight is determined by both intrauterine growth and gestation length, distinct clinical concepts with different underlying causes which have been associated with different health

[4] Adult socioeconomic status represents the cohort member's individual status. However, in cases where the cohort member was not active in the labor market, the cohort member's spouse was used as the reference person. Changes in women's labor force participation would imply changes in the reference person reflected in the socioeconomic status measure. The large changes in women's labor force participation imply that changes in the socioeconomic measure across waves could reflect the differences in spouses' socioeconomic statuses.

consequences in later life (reviewed in Huxley, Neil & Collins 2002; Victora et al. 2008). However, the present analysis is limited to the aggregate effects of intrauterine growth and gestation as they are reflected in the available measures of socio-economic attainment and adult health in the NCDS. Despite this general measure of early health risks mixing differential effects of different underlying causes, the findings are robust to alternative definitions of early child health using different combinations of birth weight and gestation. Additional measures of child health include the numbers of chronic conditions that are reported by expert health assessments at ages 7 and 16.

Measures of mother's health status and family social class at birth are included and may be interpreted as proxy measures for either inherited traits and/or unobserved parent behaviors which may affect child health and schooling. Additional parent behaviors that may reflect efforts to compensate the negative effects of early child health status are not taken into consideration. Such compensating parental investments have well known effects in offsetting the consequences of early health and are likely correlated with measures of family socioeconomic background. In the absence of accounting for such effects, the reported estimates approximate lower bounds of early health effects in the presence of such parental behaviors.

The main pathways of interest linking early health to adult attainment are the cognitive and non-cognitive skills which, jointly with educational attainment, have been shown to be important predictors of adult labor market outcomes and socioeconomic status. An aggregate measure of mathematics, reading and writing scores at age 11 is adopted to reflect student achievement. Student behaviors which may enhance learning in school are reflected in a measure of teacher behavior assessment. This measure assesses student ability to follow instructions, complete tasks and cooperate with classmates.[5]

[5] Additional descriptions of the measures adopted for the model estimation are detailed in Table 1.

Given these available measures in the NCDS, the estimated ordered probit specifies health at age 46 as a function of a set of socio-demographic measures (*X*) collected over all seven waves, prior attainment and the baseline proxies for individuals' inherited endowments of health:

$$H_{i46} = \sum_{w=0}^{7} X_{iw}\lambda_w + gS_{i42} + \beta_{46}h_{i0} + e_{i46} \qquad (5)$$

There are numerous possible relations which may be included in a model of attainment over such a wide span of the life course. However, to emphasize the effects of early life health on subsequent socioeconomic attainment and the consequences for adult socioeconomic gradients in health, the estimated model emphasizes early health and the acquisition of skills for socioeconomic attainment. The effect of lagged socioeconomic status *(g)* represents the measure of the socioeconomic gradient in health. This coefficient constitutes the main measure of interest for evaluating the consequences of early health for adult outcomes in the presence of selection bias.

## IV. Attrition in the BCS

The NCDS is a longitudinal survey that follows a sample of individuals born during a single week in 1958 in England. The study collected extensive information about child health, development, educational attainment and continues to collect measures of adult labor market activities. The baseline measure at birth has been followed by seven waves up to age 46. This cohort provides an ideal sample for assessing the scope of sample selection bias in models of socio-economic status and health over the life course. The NCDS has among the widest coverage of the life span among longitudinal studies including measures of child development. Measures of child development and health were independently reported by teachers and health professionals and have repeated measurements during childhood. The extensive set of measures include birthweight, gestation and mothers' obstetric outcomes as well as adult socioeconomic outcomes and self-assessed health status. The main variables that we consider are described in Table 1. The steep socioeconomic gradients in health that may be observed in the NCDS also present ample variation in health and socioeconomic status for testing life course hypotheses of human development. Figure 2 presents the bivariate relationship between the probability of reporting fair or poor health and individual current socioeconomic status. Gradients are evident during all periods and the slopes display a gradual steepening over time.

The NCDS has also been the subject of past efforts to identify the magnitude of sample selection bias. Plewis et al. (2004) describe the considerable efforts NCDS managers undertook to minimize survey non-response at all waves of the study and report analyses of the predictors of non-response that were used to inform refreshment sampling. Additional efforts to identify the severity of survey non-response also explicitly evaluated the magnitude of sample selection bias (Hawkes and Plewis 2006). Both studies concluded that the evidence of nonrandom non-response could only minimally affect most studies of interest using the NCDS and attribute the refreshment samples with mitigating most concerns about the effects of sample selection. However, they provide little insight into the possible severity of sample selection bias related to health that may be particularly troublesome for studies of health and socioeconomic attainment. Moreover, life course studies that examine early life events are unable to take advantage of the corrective effects from refreshment samples in the NCDS. Both concerns in life course studies require reconsidering the consequences of selection.

A closer look at the patterns of attrition across the waves in the NCDS illustrates the importance of health and socioeconomic attainment in the selective forces influencing the available data. Table 2 reports the changes in sample size and composition across the seven waves in the NCDS. The table indicates the number of respondents for each wave and the number of drop-outs that occur between waves. The survival rate is calculated as the ratio of respondents available at time *t* to the baseline sample size. The attrition rate is the ratio of the number of drop-outs between waves *t* and *t-1* to the sample of respondents in wave *t-1*. Both rates include respondents who attrited in prior waves and then returned to the sample. Such survey churning constitutes substantial shares of the sample across the seven waves, accounting for as much as 36 percent of the respondents present in Wave 7. Attrition rates are variable across waves but display the largest changes between baseline and wave 1 and when cohort members reach ages 23 and 33 in Waves 4 and 5. Both survival and attrition are plotted in Figure 1A to illustrate the constant decline in sample size over the duration of the NCDS. These high attrition rates amount to a survival rate of 52% by Wave 7.

Table 2 also reports the attrition rates for select sub-groups according to the main covariates of interest in life course studies of health and economic attainment. Corresponding trends in attrition by these covariates are plotted in Figure 1B. The differences in attrition rates across these characteristics begin to illustrate the patterns of selection over the NCDS survey that may be consequential for studying outcomes at Wave 7. While women and men display similar trends in attrition across waves, women have substantially lower attrition rates in every period except Wave 3. The differences by birthweight status are larger. Individuals born low birthweight attrite between baseline and Wave 1 at nearly triple the rate of individuals with normal birth weight. Most of this gap is attributed to a 4.95% child mortality rate in the NCDS that largely occurs between baseline and Wave 1 and disproportionately affects low birthweight cohort members. Higher mortality rates among boys throughout childhood also accounts for approximately half the gap in attrition between men and women over the first three waves. However, low birthweight cohort members who survive childhood drop out at rates that are 14, 24 and 19 percent higher than normal birthweight cohort members at ages 23, 33 and 46. Cohort members with low social class backgrounds also drop out at rates between 38% and 50% greater than individuals from the highest two social classes over all four adult waves. Comparable differences in attrition rates occur between individuals with poor measures of behavior measured in wave 2. Such sustained differences in attrition over the course of the survey result in the substantial differences in survival reported at the bottom of Table 2. These changes in the composition of the sample may severely bias estimates in models relating baseline characteristics to adult outcomes and begin to illustrate the selective forces for survey attrition which are related to early health and socioeconomic status.

Table 3 reports the probit model estimates for participation that form the basis for constructing the weighting scheme. Probits are estimated for the six waves in which the measures that enter the main model for health attainment are collected.[6] The dependent variables for these probits equal 1 if the individual participates in the survey and 0 otherwise. The specifications in Equation 3 model the probability of response as a function of measures from baseline to wave *t-1.* The flexible form for Equation 3 permits adopting different sets of variables over the course of the survey that may reflect changing individual characteristics and life experiences and their consequences for the probability of survey participation.

The table illustrates multiple dimensions of sample selection that pose a mixed set of consequences for standard estimates of life course models spanning birth to middle adulthood. The first panel illustrates the differential probabilities of participation related to sex and family structure during adulthood. While the high rate of attrition among men in Wave 1 is attributed to the differential child mortality discussed above, higher attrition among men reemerges by Wave 5. Marital status and fertility effects also show large significant associations with attrition in

[6] Although education attainment was recorded in Wave 4, it was validated with independent administrative records and this validated measure appeared in Wave 5.

Waves 6 and 7. All models include the eleven available indicators for the geographic region of birth and show additional large and significantly higher probabilities of participation for rural regions and Scotland relative to London. The clearest evidence of the selective forces of survey participation are the sustained negative effects of health. Nearly half of the low birthweight status effect may be attributed to the disproportionate rates of child mortality among this population during their first five years of life. While birthweight status does not display direct effects on participation past Wave 1 (not shown), significant correlates which may be along the pathways linking birthweight to later life health show large estimates. The count of chronic conditions at age 7, the number of absent school days due to health and adult smoking intensity and body mass index significantly associate with participation. The aggregate effects over the life course that are implied by these estimates suggest a potentially cumulative nature of selection whereby the growing risks for falling into poor health over the life course is accompanied by a similarly growing risk for survey attrition.

Socioeconomic status and the inputs for educational achievement show similarly large associations with participation, although the effects occur in different directions. Individuals from low socioeconomic background display higher probabilities of participation relative to individuals in the highest socioeconomic category (professionals). By contrast, childhood achievement favors sustained participation—test scores at ages 7 and 11 and educational attainment are both positively associated with participation. Adding further uncertainty to the net effects among these important inputs for later socioeconomic attainment, worse scores of teacher assessed behaviors (maladjustment) are associated with *lower* probabilities of remaining in the sample. In summary, significant differences in participation by sex and family structure and by health and socioeconomic status over the life course illustrates a set of selective forces occurring along the main dimensions emphasized in extant hypotheses linking health and socioeconomic attainment over the life course.

## V. Results

Tables 4 through 6 present the results for ordered probit estimates of self-reported health at age 46. The first table presents the results for an uncorrected sample in which the observations included in the analytic sample are selected by both attrition and item non-response. This case complete sample has a sample size of 1,977 and presents the basis for comparison with estimates from samples which have been corrected for attrition and item non-response.

The unconditional socioeconomic gradient in self-reported health is reported in Model 1. The effects of lagged socioeconomic status are large and show a decline across the six categories. An individual in the lowest category (1) has a probability of poor health relative that is more than twice as high relative to individuals in the highest socioeconomic category. The difference is greatest between the indicators for the two lowest classes (1 and 2), although the decline is significant and apparent over all five indicators. Model 2 demonstrates the importance of educational attainment in mediating the gradient. The attenuation and reduction in both the size of the gradient estimates and the slope over the five indicators is consistent with findings from cross-sectional studies in which respondents report both socioeconomic status and education. Model 3 confirms that the education effect is also robust to family of origin effects. In this case, the estimates from the case complete sample are consistent with Kitigawa and Hauser’s (1973) emphasis on the changes in behaviors and socioeconomic trajectories which may be attributable to education among the origins of adult socioeconomic gradients in health.

The remaining models introduce the pathways related to cognitive and non-cognitive skills and childhood health. The significant effect of cognition and its attenuation of the education effects suggests that the behaviors and abilities associated with this measure underlie the widely reported correlations between education and adult health.  The remaining measures of childhood behavior and chronic conditions at age 7 show modest effects consistent with a broader set of pathways among the origins of adult socioeconomic gradients.

Table 5 reports a comparable set of estimates from a sample which has been corrected for item non-response with multiple imputation.  The coefficients and standard errors have been corrected for both within-sample variance and the between-sample variance across the ten datasets generated by the multiple imputation approach.  The table includes the mean log likelihoods across the multiple datasets as well as the minimum and maximum values to provide an indication of the variance in the likelihoods.  The expected gain in efficiency is evident for many of the estimates across all six models.  Accounting for item non-response reduces the magnitude of the socioeconomic gradient for four of the five categories and lowers the effect of the lowest category by approximately 30 percent.  Despite the larger magnitudes of the education effects with the multiple imputation sample, education effects account for a smaller share of the socioeconomic gradient, attenuating the effects for only two of the four socioeconomic categories.

The estimates from the multiple imputation sample further suggest different adolescent pathways leading to the adult socioeconomic gradient in health.  The effect of cognition is fully attenuated.  Combined with the sustained effect of maladjustment, these changes suggest a more important role for maladjustment rather than early achievement for lifetime trajectories of socioeconomic status and health.  The modest effects of child chronic conditions at age 7 in the case complete sample are similarly attenuated.

The results for the attrition corrected sample are presented in Table 6.  There are two main results.  First, there is a further reduction in the unconditional gradient.  Model 1 shows not only a downward shift in the gradient across the five indicators, but also a modest reduction in the slope.  Figure 3 illustrates these changes in the gradient by plotting the estimated coefficients from each of the three samples.  Second, there is a reduction in the contribution of education to the gradient.  Although the estimates are of comparable magnitude to the effects from the multiple imputation sample, there is an attenuation for only one of the four coefficients for socioeconomic status.  Moreover, as in the multiple imputation sample, there is a further increase in the magnitude of the education effects.  The remaining effects are comparable to the effects from the multiple imputation sample with only modest changes in the magnitudes for coefficient estimates.

## VI.  Conclusions

This paper introduces a method for correcting the effects of sample selection bias time that may be attributed to item non-response and survey attrition.  The proposed method is implemented with a longitudinal study to examine the consequences that sample selection bias may have for estimates of socioeconomic gradients in health.  A life course model for health and socioeconomic attainment is developed that introduces cognitive, non-cognitive and early health pathways among the origins of adult socioeconomic gradients in health.  The model is then estimated with an uncorrected sample and samples reflecting corrections for item non-response alone and the combination of item non-response and survey attrition.

The evidence of selective attrition along socioeconomic and health characteristics in the NCDS are consistent with the changes in estimates across the three samples.  The reduction in both the size and slope of the gradient is consistent with selective forces over the life course by education and cognition.  Further reduction in the importance of education in the relationship between adult socioeconomic status and health suggests that many accounts which attribute adult gradients to changes in behavior associated with education may also suffer bias.  Finally, the reduction in the direct effects of early achievement and the gain in importance of childhood non-cognitive skills when sample selection bias is taken into consideration suggest a wide scope for such bias to influence interpretations of the origins of adult socioeconomic gradients in health.

## References

Adams, Peter, Michael D. Hurd, Daniel McFadden, Angela Merrill, and Tiago Ribeiro. 2003. "Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status." *Journal of Econometrics* 112: 3-56.

Almond, D. 2006. "Is the 1918 Influenza Pandemic Over? Long-term Effects of In Utero Influenza Exposure in the Post-1940 U.S. Population." *Journal of Political Economy* 114(4): 672-712.

Barker, D. J. P. 1995. "Fetal origins of coronary heart disease." *British Medical Journal* 311: 171-174.

Barker, D. J., P. D. Gluckman, K. M. Godfrey, J. E. Harding, J. A. Owens and J. S. Robinson. 1993. "Fetal nutrition and cardiovascular disease in adult life." *Lancet* 341: 938-41.

Behrman, Jere R. and Mark R. Rosenzweig. 2004. "Returns to Birth weight." *Review of Economics and Statistics* 86(2)(May): 586-601.

Black, S. E., P. J. Devereux and K. G. Salvanes. 2007. "From the cradle to the labor market? The effect of birth weight on adult outcomes." *Quarterly Journal of Economics* 122(1): 409-439.

Bleakley, H. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics* 122(1): 73-117.

Bound, J. and A. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9: 1-24.

Chadwick, E. 1842. *Report of an Enquiry into the Sanitary Conditions of the Laboring Population of Great Britain.* London: Poor Law Commission.

Clarke, Sian E et al. 2008. "Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: a cluster-randomised, double-blind, placebo-controlled trial." *The Lancet* 372: 127-138.

Conley, D., K. Strullly, and N. Bennett. 2003. "A Pound of Flesh or Just Proxy? Using Twin Differences To Estimate The Effect of Birth Weight on Life Chances." *National Bureau of Economic Research Working Paper Series*.

Contoyannis, P., A. M. Jones and N. Rice. 2004. "The Dynamics of Health in the British Household Panel Survey." *Journal of Applied Econometrics* 19: 473-503.

Cunha, F. and J. J. Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43(4).

Currie, J., and E. Moretti. 2003. "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings." *Quarterly Journal of Economics* 118: 1495 - 1532.

Dowd, Jennifer B, and Noreen Goldman. 2006. "Do biomarkers of stress mediate the relation between socioeconomic status and health?." *Journal of Epidemiology and Community Health* 60: 633-639.

Ecob, R., and G. Davey Smith. 1999. "Low Birthweight Children: Coping in School?." *Acta Paediatrica* 91: 939 - 945.

Gornick, Marian E. et al. 1996. "Effects of Race and Income on Mortality and Use of Services among Medicare Beneficiaries." *New England Journal of Medicine* 335: 791-799.

Grossman, Michael. 1972. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy* 80: 223.

Grovers, R. M. and M. P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley and Sons.

Hawkes, D. and I. Plewis. 2006. "Modelling non-response in the National Child Development Study." *Journal of the Royal Statistical Society* A 169(3): 479-491.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error," *Econometrica* 47(1): 153-162.

Huxley, R., A. Neil and R. Collins. 2002. "Unraveling the fetal origins hypothesis: is there really an inverse association between birthweight and subsequent blood pressure?" *Lancet* 360: 659-665.

Kaplan, G. A., T. E. Seeman, R. D. Cohen, L. P. Knudsen, and J. Guralnik. 1987. "Mortality among the Elderly in the Alameda County Study: Behavioral and Demographic Risk Factors." *American Journal of Public Health* 77: 307-312.

Kirby, J. B., and T. Kaneda. 2006. "Access to Health Care - Does Neighborhood Residential Instability Matter." *Journal of Health and Social Behavior* 47: 142-155.

Kitigawa, E. and P. Hauser. 1973. *Differential Mortality in the United States*. Cambridge: Harvard University Press.

Kuh, D., Y. Ben-Shlomo, J. Lynch, J. Hallqvist and C. Power. 2003. "Life course epidemiology." *Journal of Epidemiology and Community Health* 57: 778-783.

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

Marmot, M. G. et al. 1991. "Health Inequalities among British Civil Servants: The Whitehall II Study." *Lancet,* 337: 1387-1393.

Oreopoulos, P., M. Stabile, R. Walld and L. L. Roos. 2008. "Short-, Medium-, and Long-Term Consequences of Poor Infant Health: An Analysis Using Siblings and Twins." *Journal of Human Resources* 43(1): 88-138.

Palloni, A., R. G. White, and C. Milesi. 2009. The size of health selection effects. Working Paper 2009-01, Center for Demography and Ecology, University of Wisconsin-Madison.

Plewis, I., L. Calderwood, D. Hawkes and G. Nathan. 2004. "National Child Development Study and 1970 British Cohort Study Technical Report: Changes in the NCDS and BCS70 Populations and Samples over Time," London: Institute of Education.

Preston, S. H. and P. Taubman. 1994. "Socioeconomic Differences in Adult Mortality and Health Status." Pp. 279-318 in *The Demography of Aging,* eds. L. Martin and S. Preston. Washington, D.C. National Academy Press.

Raghunathan, T. E., J. M. Lepkowski, J. H. van Hoewyk and P. Solenberger. 2001. "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology* 27: 85-95.

Ross, C. E., and C.-L. Wu. 1995. "The Links between Education and Health." *American Sociological Review* 60: 719-745.

Rubin, D. B. 1976. "Inference and missing data." *Biometrika* 63: 581-592.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Shillingford, Amanda J. et al. 2008. "Inattention, Hyperactivity, and School Performance in a Population of School-Age Children With Complex Congenital Heart Disease." *Pediatrics* 121: e759-767.

Shorrocks, A. F. 1975. "The age-wealth relationship: A cross-section and cohort analysis." *Review of Economics and Statistics* 57: 155-163.

Smith, J. P. 1999. "Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status." *Journal of Economic Perspectives* 13(2): 145-66.

Smith, J. P. 2004. "Unraveling the SES-health connection." *Population and Development Review* 30: 108-132.

Stafford, M., and M. Marmot. 2003. “Neighborhood Deprivation and Health: Does it Affect Us All Equally?” *International Journal of Epidemiology* 32: 357-366.

Todd, Petra E., and Kenneth I. Wolpin. 2007. “The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps.” *Journal of Human Capital* 1: 91-136.

Victora, C. G., L. Adair, C. Fall, et al. 2008. “Maternal and child undernutrition: consequences for adult health and human capital.” *Lancet* 371: 340-357.

Woolridge, J. M. 2002. “Inverse probability weighted M-estimators for sample selection, attrition and stratification.” *Portuguese Economic Journal* 1: 117-139.

Woolridge, J. M. 2007. “Inverse probability weighted M-Estimation for General missing data problems.” *Journal of Econometrics* 14(2): 1281-1301.

**Table 1: Variable Descriptions**

*Early Childhood*

| | |
|---|---|
| SES | SES Group of mother's resident husband or partner. Six-categories from 1951 Registrar General's Classification: (1) unskilled manual, (2) semi-skilled manual, (3) skilled manual, (4) skilled non-manual, (5) managerial & technical, (6) professional. If this information was missing at birth because the mother was single or the father had no available information, we use the mother's own SES before pregnancy (about 3% of the cases at birth). For SES measures during childhood, we use the SES of mother's resident husband or partner. If this measure is missing we use the mother's SES. |
| Parent's Education | (1) Mother's age at which schooling was completed.<br>(2) Father's age at which schooling was completed. |
| Maternal Behaviors | An indicator of maternal smoking during pregnancy. |
| Health | (1) Indicator of low birth weight: less than 88 ounces (2,500 grams) at birth.<br>(2) Number of chronic conditions at age 7, reported by a medical practitioner who indicates whether or not the child exhibits each of the following conditions: general motor handicap, disfiguring condition, mental retardation, emotional maladjustment, head and neck abnormality, upper limb abnormality, lower limb abnormality, spine abnormality, respiratory system problem, alimentary system problem, urogenital system problem, heart condition, blood abnormality, skin condition, epilepsy, other central nervous system condition, or diabetes. In the analysis we use a set of indicators of whether the cohort member had 0, 1, 2, or 3 or more chronic conditions. |

*Adolescence*

| | |
|---|---|
| Cognitive skills | Standardized average of four test scores, measuring four cognitive domains: Verbal, Non-Verbal, Reading Composition and Mathematics; measured when cohort members were 11 years old. Expressed as z-score. |
| Non-cognitive skills | The natural log of a score of behavioral maladjustment. This score is equivalent to the sum of twelve items representing different aspects of behavioral deviance reported by teachers for cohort members at age 11. |
| Health | Number of chronic conditions at age 16. In addition to the conditions reported at age 7, a medical practitioner indicated whether the child had any eye, hearing or speech condition. |
| SES | (1) Family Rents Residence Age 11<br>(2) Crowding at Home Age 11 |

*Adult Socioeconomic and Health Attainment*

| | |
|---|---|
| Educational attainment | Set of indicators of highest completed education. Indicators represent a five point scale that is derived from the numbers of passed O-level and A-level exams, completed higher education and professional certification. Information collected by schools when cohort members were 20 years old. |
| SES | Classification of cohort member's own SES , measured at ages 33, 41 and 46. Same 6-category classification as parental SES . When predicting adult SES we treat this variable as an ordered one and use ordered logit models to assess the effects of covariates. |
| Health | Individuals' own self-reported health status at ages 33, 41 and 46. To simplify analyses we grouped the four-point original scale into two categories: poor or fair (=1); and good or excellent (=0). When predicting adult health status we treat this variable as categorical and estimated logistic regressions. |

**Table 2. Sample Size, Drop-outs and Attrition by Wave and Select Covariates**

| | Full Sample | | | | Sex | | Birthweight | | Class at Birth[3] | | Maladjustment[4] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *Women* | *Men* | *Normal* | *Low* | *High* | *Low* | *Low* | *High* |
| *Wave (age)* | *Respondents* | *Survival(%)* [1] | *Dropouts* | *Attrition(%)* [2] | *Attrition(%)* | | *Attrition(%)* | | *Attrition(%)* | | *Attrition(%)* | |
| Baseline | 17,415 | | | | | | | | | | | |
| 1 (7) | 14,888 | 85.5 | 2,527 | 14.5 | 13.8 | 15.2 | 12.9 | 36.4 | 14.8 | 15.0 | - | - |
| 2 (11) | 14,573 | 83.7 | 1,231 | 8.3 | 8.3 | 8.2 | 8.2 | 9.4 | 8.6 | 7.9 | - | - |
| 3 (16) | 13,701 | 78.7 | 1,676 | 11.5 | 11.6 | 11.4 | 11.5 | 11.5 | 12.2 | 11.3 | 10.7 | 11.9 |
| 4 (23) | 11,889 | 68.3 | 2,876 | 21.0 | 18.7 | 23.2 | 20.8 | 23.7 | 17.8 | 24.6 | 18.1 | 25.5 |
| 5 (33) | 10,894 | 62.6 | 2,530 | 21.3 | 18.8 | 23.8 | 21.0 | 26.1 | 17.8 | 25.4 | 19.1 | 26.9 |
| 6 (42) | 10,830 | 62.2 | 1,480 | 13.6 | 12.1 | 15.1 | 13.6 | 13.2 | 10.1 | 15.1 | 12.3 | 16.1 |
| 7 (46) | 9,057 | 52.0 | 2,210 | 20.4 | 19.2 | 21.7 | 20.2 | 24.1 | 16.4 | 24.0 | 18.4 | 26.9 |
| | | | | *Survival* | 55.2 | 49.1 | 53.2 | 36.1 | 57.3 | 47.0 | 61.7 | 49.6 |

Notes: Survival and attrition rates include respondents who attrited and returned in preceding waves. [1] Survival is the ratio of respondents at wave t to the baseline sample. [2] Attrition is the ratio of dropouts between waves t and t-1 to the number of respondents in wave t-1. [3] High class represents aggregate social classes 5 and 6; low class represents aggregate classes 1 and 2. [4] Maladjustment is measured in Wave 2. Low maladjustment is defined by the bottom 75 percentile of maladujstment scores.

**Table 3. Probit Models for Participation/Attrition by Wave.**

| | Wave 1<br>(Age 7) | Wave 2<br>(Age 11) | Wave 3<br>(Age 16) | Wave 5<br>(Age 33) | Wave 6<br>(Age 41) | Wave 7<br>(Age 46) |
|---|---|---|---|---|---|---|
| *Individual Characteristics* | | | | | | |
| Male | 0.880*** | 0.988 | 1.043 | 0.765*** | 0.782*** | 0.804*** |
| | (0.041) | (0.066) | (0.071) | (0.051) | (0.054) | (0.047) |
| Lag Never Married | | | | 0.830*** | 0.837** | |
| | | | | (0.056) | (0.068) | |
| Number Children Age 41 | | | | | | 1.170*** |
| | | | | | | (0.053) |
| *Health* | | | | | | |
| LBW | 0.259*** | | | | | |
| | (0.018) | | | | | |
| Chronic Conditions Age 7 | | | 0.936* | | | |
| | | | (0.033) | | | |
| Missed School due Health Age 16 | | | | 0.898** | | |
| | | | | (0.042) | | |
| Top Quartile BMI Age 41 | | | | | | 0.879** |
| | | | | | | (0.056) |
| Lag Number cigarettes | | | | | 0.984*** | 0.982*** |
| | | | | | (0.003) | (0.003) |
| *Socioeconomic Status* | | | | | | |
| Lag Social Class 1 [Class 6][1] | 1.473*** | 1.841*** | 1.382* | | | |
| | (0.178) | (0.374) | (0.269) | | | |
| Lag Social Class 2 | 1.507*** | 1.650*** | 1.405** | | | |
| | (0.174) | (0.268) | (0.215) | | | |
| Lag Social Class 3 | 1.493*** | 1.255 | 1.271* | | | |
| | (0.152) | (0.182) | (0.174) | | | |
| Lag Social Class 4 | 1.632*** | 1.199 | 1.625*** | | | |
| | (0.197) | (0.203) | (0.281) | | | |
| Lag Social Class 5 | 1.675*** | 1.168 | 1.326* | | | |
| | (0.197) | (0.186) | (0.197) | | | |
| Lag Currently Unemployed | | | | | 0.749*** | 0.652*** |
| | | | | | (0.060) | (0.050) |
| *Achievement and Behavior* | | | | | | |
| Mean Achievement Age 7 | | 1.156*** | | | | |
| | | (0.045) | | | | |
| Mean Achivement Age 11 | | | | 1.164*** | 1.331*** | 1.417*** |
| | | | | (0.050) | (0.055) | (0.052) |
| Maladjustment Age 11 | | | 0.918** | 0.904*** | 0.952 | 0.914*** |
| | | | (0.031) | (0.031) | (0.033) | (0.028) |
| Maladjustment Age 16 | | | | 0.979** | | |
| | | | | (0.009) | | |
| Education – Any O-Levels [None] | | | | 1.120 | | 1.219* |
| | | | | (0.122) | | (0.132) |
| Education – A-Levels or Higher | | | | 1.315*** | | 1.259** |
| | | | | (0.136) | | (0.125) |
| *Parents' Measures* | | | | | | |
| Number Antenatal Visits | 1.024*** | | | | | |
| | (0.005) | | | | | |
| Maternal Health 1 Risk [None] | 0.823*** | | | | | |
| | (0.042) | | | | | |
| Maternal Health 2-4 Risks | 0.604*** | | | | | |
| | (0.093) | | | | | |
| N | 16,205 | 13,277 | 10,239 | 6,677 | 9,030 | 8,285 |

[1] Social Class 6 is the highest class (professional).

Notes: Exponentiated coefficients. Standard errors in parentheses. Reference categories in brackets. All models include indicators for the eleven coded regrions of birth. Waves 1-3 include separate measures for father's and mother's completed education. * p<0.10, ** p<0.05, *** p<0.01

**Table 4. Ordered Probit Models for Self-Reported Health at Age 46 (1=Excelllent, 4=Poor). Case Complete Sample**

| | M1 | M2 | M3 | M4 | M5 | M6 |
|---|---|---|---|---|---|---|
| Class: Unskilled Age 41 [Professional] | 2.361$^{***}$ | 2.088$^{***}$ | 2.077$^{***}$ | 1.959$^{***}$ | 1.965$^{***}$ | 1.944$^{***}$ |
| | (0.465) | (0.423) | (0.422) | (0.401) | (0.402) | (0.399) |
| Class: Semi-skilled Manual Age 41 | 1.341$^{**}$ | 1.184 | 1.175 | 1.125 | 1.141 | 1.148 |
| | (0.172) | (0.162) | (0.161) | (0.156) | (0.158) | (0.160) |
| Class: Skilled Manual Age 41 | 1.266$^{**}$ | 1.147 | 1.141 | 1.099 | 1.107 | 1.111 |
| | (0.127) | (0.124) | (0.124) | (0.121) | (0.122) | (0.122) |
| Class: Skilled Non-manual Age 41 | 1.294$^{**}$ | 1.199 | 1.194 | 1.187 | 1.206 | 1.212 |
| | (0.152) | (0.146) | (0.146) | (0.145) | (0.148) | (0.149) |
| Class: Managerial and Technical Age 41 | 1.139 | 1.092 | 1.087 | 1.081 | 1.099 | 1.109 |
| | (0.111) | (0.109) | (0.108) | (0.108) | (0.110) | (0.111) |
| Education: Less Than 5 O-Levels [None] | | 0.905 | 0.915 | 0.965 | 0.969 | 0.975 |
| | | (0.081) | (0.082) | (0.089) | (0.089) | (0.091) |
| Education: 5+ O-Levels Passed | | 0.961 | 0.970 | 1.018 | 1.025 | 1.035 |
| | | (0.080) | (0.081) | (0.087) | (0.088) | (0.089) |
| Education: Any A-levels or Certificate | | 0.867$^{*}$ | 0.878$^{*}$ | 0.941 | 0.949 | 0.958 |
| | | (0.064) | (0.065) | (0.075) | (0.075) | (0.076) |
| Education: University Degree Earned | | 0.791$^{**}$ | 0.814$^{**}$ | 0.922 | 0.936 | 0.943 |
| | | (0.077) | (0.082) | (0.104) | (0.106) | (0.107) |
| Class: Unskilled Birth [Professional] | | | 1.229 | 1.203 | 1.221 | 1.237 |
| | | | (0.189) | (0.185) | (0.188) | (0.191) |
| Class: Semi-skilled Manual Birth | | | 1.216 | 1.185 | 1.182 | 1.195 |
| | | | (0.162) | (0.158) | (0.158) | (0.160) |
| Class: Skilled Manual Birth | | | 1.127 | 1.110 | 1.117 | 1.128 |
| | | | (0.132) | (0.130) | (0.131) | (0.132) |
| Class: Skilled Non-manual Birth | | | 1.109 | 1.104 | 1.110 | 1.119 |
| | | | (0.146) | (0.145) | (0.146) | (0.147) |
| Class: Managerial and Technical Birth | | | 1.170 | 1.172 | 1.176 | 1.181 |
| | | | (0.147) | (0.147) | (0.148) | (0.149) |
| Cognition Age 11 | | | | 0.910$^{**}$ | 0.925$^{**}$ | 0.922$^{**}$ |
| | | | | (0.035) | (0.037) | (0.036) |
| Ln Maladjustment Age 11 | | | | | 1.052$^{*}$ | 1.052$^{*}$ |
| | | | | | (0.028) | (0.028) |
| Chronic 1 Cond Age 16 [0 Conditions] | | | | | | 1.014 |
| | | | | | | (0.083) |
| Chronic 2+ Cond Age 16 | | | | | | 1.082 |
| | | | | | | (0.180) |
| Chronic 1 Cond Age 7 [0 Conditions] | | | | | | 1.104$^{*}$ |
| | | | | | | (0.066) |
| Chronic 2 Cond Age 7 | | | | | | 0.912 |
| | | | | | | (0.086) |
| Chronic 3+ Cond Age 7 | | | | | | 1.011 |
| | | | | | | (0.150) |
| Low Birth Weight | | | | | | 0.921 |
| | | | | | | (0.133) |
| N | 1,977 | 1,977 | 1,977 | 1,977 | 1,977 | 1,977 |
| Log Likelihood | -2,221 | -2,217 | -2,216 | -2,213 | -2,211 | -2,208 |

Exponentiated coefficients. Standard errors in parentheses. Reference categories in brackets. Sample of men only.
$^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

**Table 5. Ordered Probit Models for Self-Reported Health at Age 46 (1=Excelllent, 4=Poor). Multiple Imputation.**

| | M1 | M2 | M3 | M4 | M5 | M6 |
|---|---|---|---|---|---|---|
| Class: Unskilled Age 41 [Professional] | 1.642$^{***}$ | 1.452$^{**}$ | 1.433$^{**}$ | 1.423$^{**}$ | 1.420$^{**}$ | 1.397$^{**}$ |
| | (0.222) | (0.207) | (0.203) | (0.203) | (0.202) | (0.199) |
| Class: Semi-skilled Manual Age 41 | 1.477$^{***}$ | 1.302$^{***}$ | 1.287$^{***}$ | 1.281$^{**}$ | 1.287$^{**}$ | 1.279$^{**}$ |
| | (0.131) | (0.126) | (0.124) | (0.124) | (0.125) | (0.124) |
| Class: Skilled Manual Age 41 | 1.247$^{***}$ | 1.138 | 1.129 | 1.124 | 1.128 | 1.124 |
| | (0.090) | (0.090) | (0.089) | (0.090) | (0.090) | (0.090) |
| Class: Skilled Non-manual Age 41 | 1.206$^{**}$ | 1.130 | 1.125 | 1.123 | 1.137 | 1.131 |
| | (0.100) | (0.101) | (0.099) | (0.099) | (0.101) | (0.100) |
| Class: Managerial and Technical Age 41 | 1.027 | 0.995 | 0.990 | 0.990 | 1.002 | 1.002 |
| | (0.073) | (0.074) | (0.073) | (0.073) | (0.074) | (0.074) |
| Education: Less Than 5 O-Levels [None] | | 0.831$^{***}$ | 0.839$^{***}$ | 0.843$^{***}$ | 0.848$^{***}$ | 0.851$^{***}$ |
| | | (0.049) | (0.049) | (0.050) | (0.050) | (0.051) |
| Education: 5+ O-Levels Passed | | 0.833$^{***}$ | 0.841$^{***}$ | 0.845$^{***}$ | 0.855$^{***}$ | 0.858$^{**}$ |
| | | (0.049) | (0.049) | (0.050) | (0.051) | (0.051) |
| Education: Any A-levels or Certificate | | 0.796$^{***}$ | 0.806$^{***}$ | 0.811$^{***}$ | 0.820$^{***}$ | 0.822$^{***}$ |
| | | (0.042) | (0.042) | (0.045) | (0.045) | (0.045) |
| Education: University Degree Earned | | 0.763$^{***}$ | 0.782$^{***}$ | 0.791$^{***}$ | 0.804$^{***}$ | 0.804$^{***}$ |
| | | (0.056) | (0.058) | (0.063) | (0.064) | (0.064) |
| Class: Unskilled Birth [Professional] | | | 1.185 | 1.183 | 1.186 | 1.184 |
| | | | (0.123) | (0.123) | (0.124) | (0.124) |
| Class: Semi-skilled Manual Birth | | | 1.202$^{*}$ | 1.199$^{*}$ | 1.204$^{*}$ | 1.207$^{*}$ |
| | | | (0.115) | (0.115) | (0.116) | (0.116) |
| Class: Skilled Manual Birth | | | 1.079 | 1.078 | 1.085 | 1.088 |
| | | | (0.091) | (0.091) | (0.092) | (0.092) |
| Class: Skilled Non-manual Birth | | | 1.092 | 1.092 | 1.101 | 1.103 |
| | | | (0.104) | (0.105) | (0.105) | (0.106) |
| Class: Managerial and Technical Birth | | | 1.132 | 1.133 | 1.138 | 1.137 |
| | | | (0.104) | (0.104) | (0.104) | (0.104) |
| Cognition Age 11 | | | | 0.990 | 1.012 | 1.011 |
| | | | | (0.024) | (0.026) | (0.026) |
| Ln Maladjustment Age 11 | | | | | 1.064$^{***}$ | 1.063$^{***}$ |
| | | | | | (0.021) | (0.021) |
| Chronic 1 Cond Age 16 [0 Conditions] | | | | | | 1.080 |
| | | | | | | (0.068) |
| Chronic 2+ Cond Age 16 | | | | | | 1.066 |
| | | | | | | (0.134) |
| Chronic 1 Cond Age 7 [0 Conditions] | | | | | | 1.075 |
| | | | | | | (0.047) |
| Chronic 2 Cond Age 7 | | | | | | 1.044 |
| | | | | | | (0.073) |
| Chronic 3+ Cond Age 7 | | | | | | 1.057 |
| | | | | | | (0.107) |
| Low Birth Weight | | | | | | 0.998 |
| | | | | | | (0.095) |
| Observations | 4,075 | 4,075 | 4,075 | 4,075 | 4,075 | 4,075 |
| Log Likelihood | -4,603 | -4,590 | -4,586 | -4,586 | -4,581 | -4,578 |
| Minimum | -4,601 | -4,589 | -4,585 | -4,585 | -4,577 | -4,574 |
| Maximum | -4,605 | -4,592 | -4,589 | -4,588 | -4,583 | -4,581 |

Exponentiated coefficients; Standard errors in parentheses. Reference categories in brackets. Sample of men from 10 multiply imputed datasets. Coefficients and standard errors are corrected for within- and between-sample variances. Model statistics are averaged across datasets.

$^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

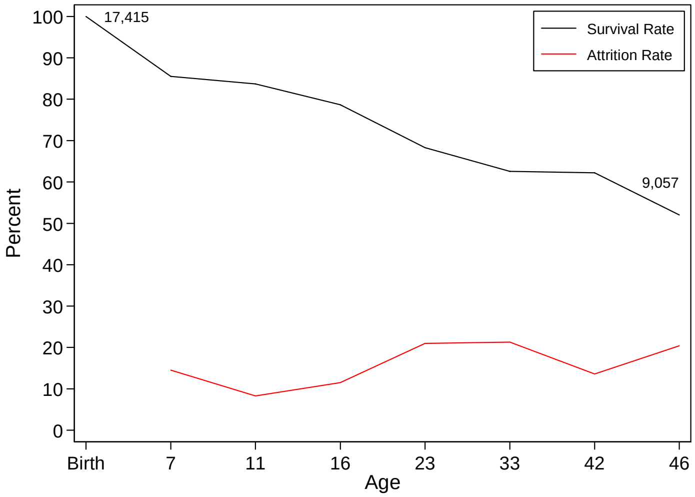
**Table 6. Ordered Probit Models for Self-Reported Health at Age 46 (1=Excelllent, 4=Poor). Attrittion Correction.**

| | M1 | M2 | M3 | M4 | M5 | M6 |
|---|---|---|---|---|---|---|
| Class: Unskilled Age 41 [Professional] | 1.558$^{***}$ | 1.394$^{**}$ | 1.373$^{**}$ | 1.374$^{**}$ | 1.371$^{*}$ | 1.346$^{*}$ |
| | (0.231) | (0.222) | (0.217) | (0.218) | (0.218) | (0.216) |
| Class: Semi-skilled Manual Age 41 | 1.509$^{***}$ | 1.342$^{***}$ | 1.328$^{***}$ | 1.329$^{***}$ | 1.333$^{***}$ | 1.322$^{***}$ |
| | (0.138) | (0.134) | (0.131) | (0.133) | (0.133) | (0.132) |
| Class: Skilled Manual Age 41 | 1.251$^{***}$ | 1.154$^{*}$ | 1.147$^{*}$ | 1.147$^{*}$ | 1.151$^{*}$ | 1.145$^{*}$ |
| | (0.090) | (0.091) | (0.091) | (0.092) | (0.092) | (0.092) |
| Class: Skilled Non-manual Age 41 | 1.195$^{**}$ | 1.131 | 1.128 | 1.129 | 1.141 | 1.134 |
| | (0.101) | (0.103) | (0.102) | (0.102) | (0.103) | (0.103) |
| Class: Managerial and Technical Age 41 | 1.004 | 0.980 | 0.976 | 0.976 | 0.987 | 0.986 |
| | (0.070) | (0.073) | (0.071) | (0.072) | (0.072) | (0.073) |
| Education: Less Than 5 O-Levels [None] | | 0.807$^{***}$ | 0.814$^{***}$ | 0.814$^{***}$ | 0.819$^{***}$ | 0.821$^{***}$ |
| | | (0.051) | (0.050) | (0.051) | (0.052) | (0.052) |
| Education: 5+ O-Levels Passed | | 0.812$^{***}$ | 0.820$^{***}$ | 0.820$^{***}$ | 0.829$^{***}$ | 0.831$^{***}$ |
| | | (0.051) | (0.051) | (0.052) | (0.052) | (0.053) |
| Education: Any A-levels or Certificate | | 0.779$^{***}$ | 0.787$^{***}$ | 0.787$^{***}$ | 0.795$^{***}$ | 0.797$^{***}$ |
| | | (0.044) | (0.044) | (0.046) | (0.046) | (0.046) |
| Education: University Degree Earned | | 0.767$^{***}$ | 0.782$^{***}$ | 0.782$^{***}$ | 0.794$^{***}$ | 0.793$^{***}$ |
| | | (0.060) | (0.062) | (0.066) | (0.067) | (0.067) |
| Class: Unskilled Birth [Professional] | | | 1.155 | 1.155 | 1.158 | 1.152 |
| | | | (0.125) | (0.125) | (0.125) | (0.124) |
| Class: Semi-skilled Manual Birth | | | 1.211$^{*}$ | 1.211$^{*}$ | 1.219$^{*}$ | 1.221$^{*}$ |
| | | | (0.121) | (0.122) | (0.122) | (0.123) |
| Class: Skilled Manual Birth | | | 1.065 | 1.065 | 1.073 | 1.075 |
| | | | (0.094) | (0.094) | (0.095) | (0.095) |
| Class: Skilled Non-manual Birth | | | 1.092 | 1.092 | 1.102 | 1.105 |
| | | | (0.109) | (0.109) | (0.110) | (0.110) |
| Class: Managerial and Technical Birth | | | 1.130 | 1.129 | 1.137 | 1.136 |
| | | | (0.106) | (0.106) | (0.107) | (0.107) |
| Cognition Age 11 | | | | 1.001 | 1.024 | 1.023 |
| | | | | (0.025) | (0.027) | (0.027) |
| Ln Maladjustment Age 11 | | | | | 1.065$^{***}$ | 1.063$^{***}$ |
| | | | | | (0.022) | (0.022) |
| Chronic 1 Cond Age 16 [0 Conditions] | | | | | | 1.110 |
| | | | | | | (0.074) |
| Chronic 2+ Cond Age 16 | | | | | | 1.109 |
| | | | | | | (0.161) |
| Chronic 1 Cond Age 7 [0 Conditions] | | | | | | 1.072 |
| | | | | | | (0.050) |
| Chronic 2 Cond Age 7 | | | | | | 1.081 |
| | | | | | | (0.080) |
| Chronic 3+ Cond Age 7 | | | | | | 1.089 |
| | | | | | | (0.112) |
| Low Birth Weight | | | | | | 1.006 |
| | | | | | | (0.095) |
| Observations | 4,075 | 4,075 | 4,075 | 4,075 | 4,075 | 4,075 |
| Log Likelihood | -4,579 | -4,565 | -4,561 | -4,561 | -4,555 | -4,551 |
| Minimum | -4,577 | -4,562 | -4,559 | -4,559 | -4,551 | -4,546 |
| Maximum | -4,581 | -4,567 | -4,563 | -4,563 | -4,559 | -4,554 |

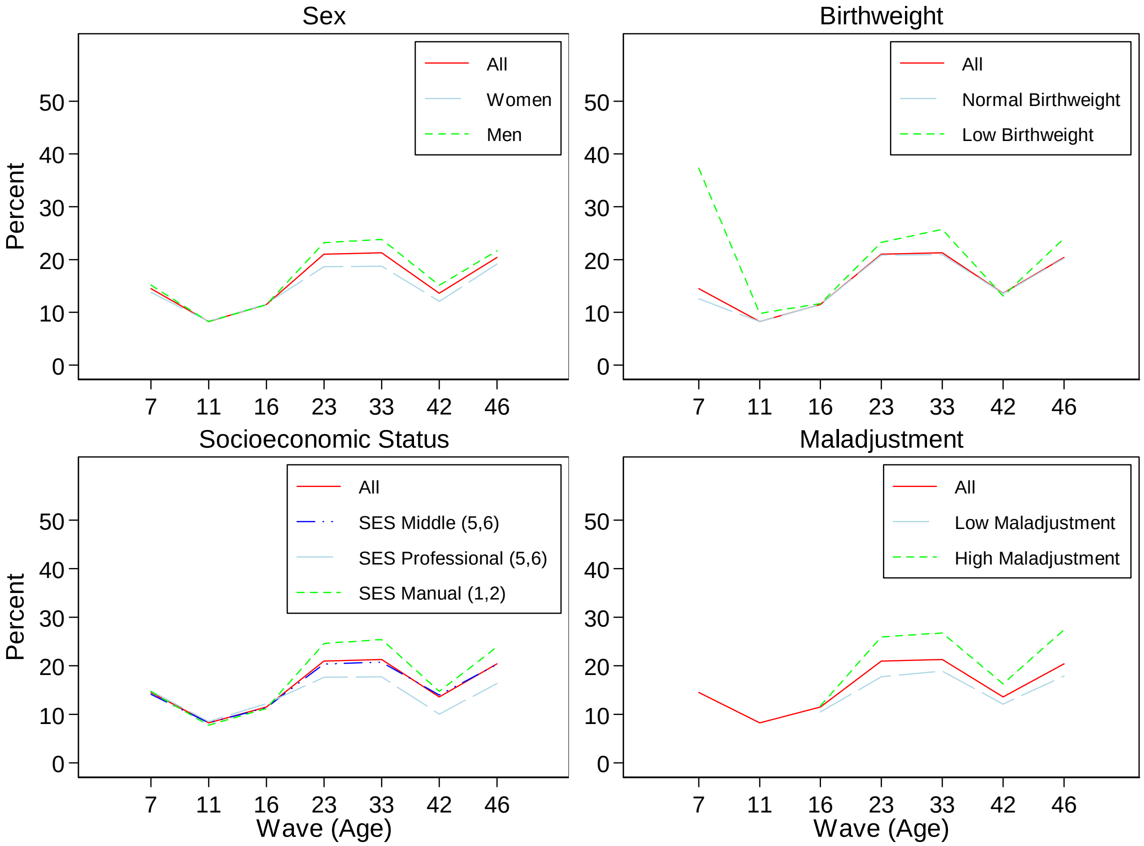
Exponentiated coefficients; standard errors in parentheses. Reference categories in brackets. Sample of men from 10 multiply imputed datasets. Coefficients and standard errors are corrected for within- and between-sample variances. Model statistics are averaged across datasets.

$^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

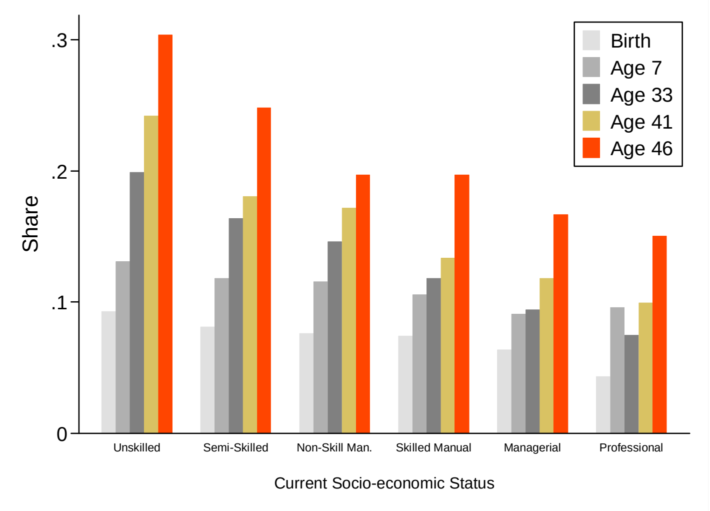
## Figure 1A. Survival and Attrition in the NCDS



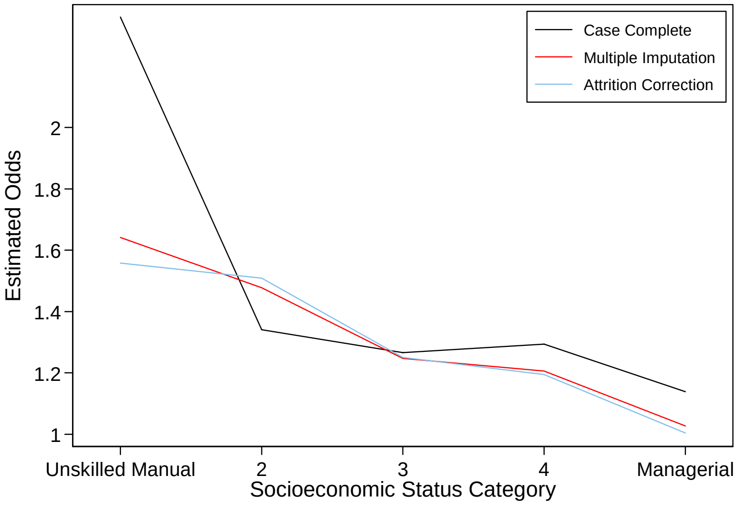
## Figure 1B. Wave Attrition by Select Covariatess



## Figure 2. Socioeconomic Gradients in Health Over Time in the NCDS



## Figure 3. Selection Bias and the Socioeconomic Gradient in Health



Note: Reference category is Professional.

**Appendix: Inverse Probability Weighting Estimator**

The IPW estimator (Woolridge 2007) enables correcting sample selection bias that is due to attrition. The estimator is implemented by first estimating the probability of participation for each wave. Probits for each wave are estimated using the sample of individuals who participate in the prior wave of the survey. The set of independent variables for each probit includes select lagged values of the current available independent variables as well as lagged participation. Incorporating lagged measures and prior participation contributes to relaxing the severity of the assumption that the selection process may be explained by observable individual characteristics. The inverse of the predicted probabilities from the set of equations described by Equation 2 are then used to weight observations during estimation. Woolridge (2007) demonstrates the consistency of the resulting IPW estimator and describes its asymptotic properties. This weighting is analagous to probability weights which are used in stratified samples. The predicted probabilities are applied in the maximum likelihood estimation of the ordered probit by the following:

$$Log(L) = \sum_{i=1}^{N}\sum_{t=0}^{T}\left( {r_{it}}/{\hat{p}_{it}} \right) \log L_{it} \qquad (6)$$

An aggregate probability weight is constructed cumulatively with each wave's probability weight in order to account for multiple waves with changes in both the probabilities of attrition and sets of conditioning variables. In this case, if $\hat{\pi}_{it}$ represents the fitted participation probability for wave *t,* the predicated probability for any period *t* may be defined by:

$$\hat{p}_{it} = \hat{\pi}_{i1} \ldots \hat{\pi}_{it} \qquad (7)$$

Woolridge (2002) shows that the resulting standard errors lead to conservative inference which does not require correcting the change in variance due to the weighting.

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: rwhite@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu