Department of Quantitative Social Science

Missing ordinal covariates with informative selection

Alfonso Miranda
Sophia Rabe-Hesketh

DoQSS Working Paper No. 10-16
November 2010

IOE LONDON
Leading education and social research
Institute of Education
University of London

DISCLAIMER

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science. Institute of Education, University of London. 20 Bedford way, London WC1H 0AL, UK.

Missing ordinal covariates with informative selection

Alfonso Miranda[*], **Sophia Rabe-Hesketh**[†‡]

Abstract. This paper considers the problem of parameter estimation in a model for a continuous response variable y when an important ordinal explanatory variable x is missing for a large proportion of the sample. Non-missingness of x, or sample selection, is correlated with the response variable and/or with the unobserved values the ordinal explanatory variable takes when missing. We suggest solving the endogenous selection, or 'not missing at random' (NMAR), problem by modelling the informative selection mechanism, the ordinal explanatory variable, and the response variable together. The use of the method is illustrated by re-examining the problem of the ethnic gap in school achievement at age 16 in England using linked data from the National Pupil database (NPD), the Longitudinal Study of Young People in England (LSYPE), and the Census 2001.

JEL classification: C13, C35, I21.

Keywords: Missing covariate, sample selection, latent class models, ordinal variables, NMAR.

[*]Department of Quantitative Social Science, Institute of Education, University of London. 20 Bedford Way, London WC1H 0AL, UK. E-mail: A.Miranda@ioe.ac.uk

[†]Graduate School of Education and Graduate Group in Biostatistics, University of California, Berkeley, USA. Institute of Education, University of London, London, UK. E-mail: sophiarh@berkeley.edu

[‡]We are grateful to Lorraine Dearden, John MacDonald, and Anna Vignoles for useful comments. This research was supported by ESRC grant RES-576- 25-0014 under the ADMIN node of the National Centre for Research Methods at the Institute of Education

1 Introduction

Applied researchers often find themselves fitting a model when important explanatory variables are missing for a substantial proportion of the sample. An important example, which motivates this paper, is when administrative data are merged with survey data to obtain covariate information for the small subset of individuals who were included in the survey. We consider the situation where one explanatory variable X has missing values, whereas the other explanatory variables Z and the response variable Y are always observed. A common practice is to discard units with missing data (sometimes referred to as list-wise deletion) and perform complete-case analysis. This approach is, in general, problematic as consistent estimators are obtained only if the probability of selection (non-deletion from the sample) does not depend on the response variable given the explanatory variables — note that consistent estimates are obtained if selection depends on X or Z (Griliches et al. 1978, Little 1992; Little and Rubin 2002, p.43; Wooldridge 2002, p.556); so, the condition for consistency is $\Pr(S|Y,X,Z) = \Pr(S|X,Z)$. Even when this condition is satisfied, complete case analysis is inefficient if a large portion of the data are missing, the problem considered in this paper.

Under the 'missing at random' (MAR) assumption that $\Pr(S|Y,X,Z) = P(S|Y,Z)$, complete-case analysis can yield consistent estimates if a weighted version of the estimation method is used with weights given by the inverse of the probability of selection, $P(S|Y,Z)$. Estimates of such inverse probability weights are usually part of survey datasets where they are referred to as design weights (to account for differential probabilities of being included in the sample) adjusted for non-response. The idea of 'response propensity weighting' is discussed by Little (1988). Robins et al. (1995) propose more efficient, 'doubly robust' estimators that also make use of cases with missing data. An excellent overview and intuitive explanation of these methods is given by Carpenter et al. (2006). Wooldridge (2007) considers inverse probability weighted M-estimation under a general missing data scheme, including

the case when the predictors of selection are not always observed.

A commonly used alternative to complete-case analysis is multiple imputation (Rubin 2002, Little and Rubin 2002, Schafer 2002). Here, the missing data are filled in by sampling from the estimated regression model $\widehat{\Pr}(X|Y,Z)$, and this is done multiple times, yielding several imputed datasets. Each imputed dataset is then analyzed using conventional methods, and the estimates are averaged across datasets. Squared standard errors are estimated as the means of the 'within-imputation' squared standard errors plus the 'between-imputation' variances of the estimates. This method makes use of all available data and is therefore more efficient than complete-case analysis. Unlike complete-case analysis, multiple imputation does not yield consistent estimates if selection depends on X given Y and Z (because the imputation model is then inconsistently estimated), whereas it no longer assumes that selection is independent of Y given Z (since Y is no longer deleted when X is missing). See Carpenter et al. (2006) for a comparison of multiple imputation and inverse probability weighting methods.

A similar approach to multiple imputation is maximum likelihood estimation of a joint model for Y and X, with missing values of X integrated out. For instance, Little and Schluchter (1985) developed an EM algorithm for maximum likelihood estimation of a model with missing continuous and categorical covariates. Bayesian estimation of a joint model for Y and X can also be accomplished relatively easily using Markov chain Monte Carlo methods where missing values of X are sampled from their posterior distribution along with the model parameters (e.g., Ibrahim et al. 2002). These likelihood and Bayesian approaches require specification of the distribution of X given Z. In the case of categorical X, loglinear models are often specified for the cell counts and in the case of continuous X a parametric distribution, such as normality is typically assumed. Little and Schluchter (1985) combine these specifications for the case of both continuous and categorical X. See Horton and Laird (1998) for a useful review of these approaches. For the case when Z is categorical with few categories, Chen (2004) suggests a nonpara-

metric approach even when X includes continuous covariates. Zhao (2009) suggests a piecewise nonparametric model for X given Z when both X and Z are continuous. All these methods make the same missing at random (MAR) assumption as multiple imputation (Little 1992, Ibrahim et al. 2005).

Vach and Blettner (1995) consider sensitivity of maximum likelihood estimation under the MAR assumption to violations of MAR. Rotnitzky and Robins (1995) extend the weighted estimating equation approach to account for nonignorable nonresponse in the covariates or the outcomes. The MAR assumption can also be relaxed by specifying a joint model for Y, X, and S. Lipsitz et al. (1999), Stubbendick and Ibrahim (2003) and others specify a selection model $\Pr(S|Y, X, Z)$, analogously to the approach by Hausman and Wise (1979) and Diggle and Kenward (1994) for missing Y. Huang et al. (2005) estimate a Bayesian version of such selection models using a Gibbs sampler.

In this paper, we handle violation of the MAR assumption by allowing the residuals for different models to be correlated through shared random effects, similar to the models by Wu and Carroll (1988) for missing Y. Our models also resemble the models for sample selection and endogenous covariates introduced by Heckman (1979). Specifically, we consider a continuous response variable Y and an ordinal explanatory variable X which is subject to missing data. Our model for Y, X, and S relaxes the MAR assumption by allowing dependence between S and X and between S and Y, controlling for Z, through a shared random effects approach.

In this paper we consider the problem of linking administrative data, where Y and Z are observed for an entire population, to survey data where X is observed only for a small subset of individuals. For the case of several surveys, Gelman et al. (1998) suggested multiple imputation for this problem with separate imputation models for different surveys linked within a hierarchical model. Jackson et al. (2009) handled the same kind of problem using Bayesian graphical models that made the MAR assumption. They found estimation of the model computationally infeasible due to the large

dataset and used a 2-stage imputation and regression approach.

We use our model to re-examine the *ethnic group gaps* in the General Certificate of Secondary Education (GCSE) test scores at age 16 in England using the National Pupil database (NPD) and the 2001 census data (which give information on the entire population of pupils in state maintained schools), linked to survey data from the Longitudinal Study of Young People in England (LSYPE). A key ordinal covariate, mother's education, is only available for pupils sampled into the LSYPE whose mothers responded to the relevant question. Our method allows us to estimate the ethnic group gaps, net of differences in background characteristics including mother's education, while exploiting GCSE and ethnicity data on the entire population. Our results show that once mother's education is controlled for, the estimated gap between Black Caribbean pupils and the White British majority is even larger than suggested by previous studies. Other minorities also do *better* in relation to the White British majority.

2 The model

Let x_i be an ordinal variable with G categories $g = \{1, \ldots, G\}$. The model for the continuous response is

$$y_i = \begin{cases} \sum_{g=1}^{G} \beta_g 1(x_i = g) + \mathbf{w}_i'\boldsymbol{\theta} + \epsilon_{yi} & \text{if } x_i \text{ is observed} \\ \eta_{1i} + \mathbf{w}_i'\boldsymbol{\theta} + \epsilon_{yi} & \text{otherwise} \end{cases} \quad (1)$$

where $1(x_i = g)$ is a dummy variable for the gth value of the ordinal variable x_i with regression coefficient β_g, $\mathbf{w}_i$ are other explanatory variables with regression coefficients $\boldsymbol{\theta}$ and η_{1i} is a discrete latent variable equal to the appropriate coefficient β_g, with g unknown because x_i is missing.

The model for the covariate x_i is an ordinal probit model that can be written as a linear model for the latent response x_i^*,

$$x_i^* = \mathbf{z}_i'\boldsymbol{\gamma} + \epsilon_{xi}, \quad (2)$$

where $\mathbf{z}_i$ are explanatory variables with regression coefficients $\boldsymbol{\gamma}$, $x_i = g$ if $\kappa_{g-1} \leq x_i^* < \kappa_g$ and κ_g are threshold or cut-point parameters with $\kappa_0 = -\infty$ and $\kappa_G = \infty$. The conditional probabilities that $\eta_{1i} = \beta_g$ are set equal to the conditional probabilities that $x_i = g$.

The selection process is a probit model with the latent response modelled as

$$S_i^* = \mathbf{r}_i'\boldsymbol{\alpha} + \epsilon_{si} \tag{3}$$

where $\mathbf{r}_i$ are explanatory variables with regression coefficients $\boldsymbol{\alpha}$ and $S_i = 1(S_i^* > 0)$.

To allow for endogenous selection, the error term ϵ_{Si} in the selection model may be correlated with the error terms ϵ_{xi} and ϵ_{yi}. The correlations are induced by two latent variables, η_{2i} and η_{3i},

$$\begin{aligned} \epsilon_{yi} &= \lambda_1\eta_{2i} + u_{yi} \\ \epsilon_{xi} &= \eta_{3i} + u_{xi} \\ \epsilon_{Si} &= \lambda_2\eta_{2i} + \lambda_3\eta_{3i} + u_{Si}, \end{aligned} \tag{4}$$

where λ_1, λ_2 and λ_3 are parameters. We assume that η_{2i} and η_{3i} are i.i.d standard normal. We further assume that u_{yi}, u_{xi} and u_{Si} are normally distributed with zero mean and a small, fixed variance such as $\sigma^2 = 0.04$.

In (4), η_{2i} induces a correlation between ϵ_{yi} and ϵ_{Si}, given by

$$\text{Cor}(\epsilon_{yi}, \epsilon_{Si}) = \frac{\lambda_1\lambda_2}{\sqrt{(\lambda_1^2 + \sigma^2)(\lambda_2^2 + \lambda_3^2 + \sigma^2)}},$$

and η_{3i} induces a correlation between ϵ_{xi} and ϵ_{Si}, given by

$$\text{Cor}(\epsilon_{xi}, \epsilon_{Si}) = \frac{\lambda_3}{\sqrt{(1 + \sigma^2)(\lambda_2^2 + \lambda_3^2 + \sigma^2)}}.$$

This parameterization is chosen to obtain zero correlation between ϵ_{yi} and ϵ_{xi} while not imposing any unnecessary constraints on the other two correlations or the variance of ϵ_{yi}.[1] This assumption is motivated by our empirical

[1] When one of the pairwise correlations between three random variables is set to zero, the sum of squares of the other two cannot be larger than 1 (e.g., Takeuchi et al. 1982, p.85).

application in section 3 where we calculate the ethnic gaps in test scores in England at age 16 and mother's education plays the role of the missing ordinal covariate. In this setting, one cannot think of a causal relationship between ethnicity and pupil's achievement and, as a consequence, once observable characteristics are controlled for, $\text{Cor}(\epsilon_{yi},\epsilon_{xi}) = 0$ is the natural assumption. The model, however, can be easily modified to allow $\text{Cor}(\epsilon_{yi},\epsilon_{xi}) \neq 0$.

The conditional normal density of y_i is denoted $\phi_{x_i\text{o}}(y_i|x_i,\eta_{2i})$ when x_i is observed and $\phi_{x_i\overline{\text{o}}}(y_i|\eta_{1i},\eta_{2i})$ when x_i is not observed. As discussed earlier, we assume that u_{xi} is normally distributed with zero mean and fixed variance σ^2 given $\mathbf{z}_i$. Denote the response probabilities for x_i and η_{1i} by

$$\begin{aligned} P_x(g|\eta_{3i}) &= P_{\eta_1}(\beta_g|\eta_{3i}) \\ &= \Phi\left(\tfrac{\mathbf{z}_i'\boldsymbol{\gamma}+\eta_{3g}-k_g}{\sigma}\right) - \Phi\left(\tfrac{\mathbf{z}_i'\boldsymbol{\gamma}+\eta_{3g}-k_{g-1}}{\sigma}\right), \qquad g = \{1,\ldots,G\}. \end{aligned} \tag{5}$$

Finally, given η_{2i} and $\mathbf{r}_i$, u_{Si} is also normally distributed with zero mean and fixed variance σ^2. The conditional probabilities for the observed selection indicator S_i are denoted by $P_S(0|\eta_{2i},\eta_{3i})$ and $P_S(1|\eta_{2i},\eta_{3i})$, with (We have suppressed conditioning on the observed covariates in the notation)

$$P_S(1|\eta_{2i},\eta_{3i}) = \Phi\left(\frac{\mathbf{r}_i'\boldsymbol{\alpha}+\lambda_2\eta_{2i}+\lambda_3\eta_{3i}}{\sigma}\right).$$

The Log-likelihood can be written as:

$$\begin{aligned} L &= \sum_{i,\ x_i\text{o},\ S_i=1} \ln\left\{\iint P_S(1|\eta_{2i},\eta_{3i})P_x(x_i|\eta_{3i})\phi_{x_i\text{o}}\left(y_i|x_i,\eta_{2i}\right)d\eta_{2i}d\eta_{3i}\right\} \\ &+ \sum_{i,\ x_i\overline{\text{o}},\ S_i=0} \ln\left\{\iint P_S(0|\eta_{2i},\eta_{3i})\left[\sum_{g=1}^{G} P_{\eta_1}\left(\beta_g|\eta_{3i}\right)\phi_{x_i\overline{\text{o}}}\left(y_i|\beta_g,\eta_{2i}\right)\right]d\eta_{2i}d\eta_{3i}\right\} \end{aligned}$$

For the summations, $x_i\overline{\text{o}}$ means that x_i was not observed, whereas $x_i\text{o}$ means that x_i was observed. Figure 1 gives a graphical representation of the model. Circles represent unobserved, or latent, variables whereas rectangles represent observed variables. Arrows connecting latent and/or observed variables represent linear and non-linear relationships. Coefficients are written alongside the relevant arrow. Subfigure (a) depicts the model when the ordinal

covariate x_i is missing while subfigure (b) depicts the model when the ordinal covariate x_i is observed.

2.1 S not always observed: Merging administrative and survey data

An important application where a large proportion of the sample will have missing covariates is when administrative data are merged with survey data. The administrative data contains the dependent variable of interest, such as employees' wages, or students' examination results, whereas the survey data provides detailed background information, such as parent's education, for just a subset of the sample in the administrative data. We assume that the individuals sampled into the survey represent a random sample (after conditioning on design variables) of the individuals included in the administrative data. For those not sampled into the survey, S_i is 'missing at random'. Due to survey non-response and item non-response, not everyone included in the survey provides information on the ordinal explanatory variable x_i of interest. This sample-selection process is modeled using (3). Combined with the same model for y_i as in (1), this gives rise to the same likelihood contributions as before for those individuals sampled into the survey. The following additional term is needed for those not sampled, for whom S_i is not observed (denoted $S_i\overline{\text{o}}$), and x_i is not observed:

$$\sum_{i,\ x_i\overline{\text{o}},\ S_i\overline{\text{o}}} \ln \left\{ \iint \left[\sum_{g=1}^{G} P_{\eta_1}(\beta_g|\eta_{3i}) \phi_{x_i\overline{\text{o}}}(\beta_g, \eta_{2i}) \right] d\eta_{2i} d\eta_{3i} \right\} \tag{6}$$

2.2 Parameter estimation

The model is estimated by Maximum Simulated Likelihood (MSL) (see, for instance, Train 2003). Denote by $L_i(\eta_{2i}, \eta_{3i})$ the likelihood contribution of the i-th individual conditional on the random effects η_{2i} and η_{3i}. Since η_{2i} and η_{3i} are unobserved variables, they must be integrated out to obtain the

marginal likelihood contribution based only on observed data L_i:

$$L_i = \int \int L_i(\eta_{2i}, \eta_{3i})\phi(\eta_{2i})\phi(\eta_{3i})d\eta_{2i}d\eta_{3i} \tag{7}$$

The integral in equation 7 does not have a closed form solution and, as a consequence, must be calculated numerically. We approximate the integral by:

$$L_i = \frac{1}{R}\sum_{r=1}^{R} L_i(\eta_{2i}^r, \eta_{3i}^r)\phi(\eta_{2i}^r)\phi(\eta_{3i}^r). \tag{8}$$

Two uncorrelated Halton sequences of dimension R are first drawn. Then $\eta_{2i}^1, \ldots, \eta_{2i}^R$ and $\eta_{3i}^1, \ldots, \eta_{3i}^R$ are obtained by transforming the Halton sequences using the inverse cumulative normal distribution. Halton sequences have been shown to achieve high precision with fewer draws than uniform pseudorandom sequences because they have better coverage of the $[0, 1]$ interval (for more on this topic see Train 2003). Maximum simulated likelihood is asymptotically equivalent to Maximum Likelihood as long as R grows faster than $\sqrt{N}$, where N is the sample size (Gourieroux and Monfort 1993).

Maximisation is performed using the Newton-Ramphson algorithm with analytical first derivatives of the likelihood function and an outer product of gradients (OPG) approximation of the Hessian (see, for instance, Berndt et al. 1974). Note that we have greatly simplified computation by using two random effects to induce the dependence among the three residuals. This strategy reduces the dimensionality of integration from three to two.

The likelihood appears to be flat in some areas and may not be globally concave. To avoid local maxima, it is therefore important to find good starting values. In particular, we found that the estimates of β_g can go astray, probably due to the large proportion of the sample with missing X. Our approach is to first find the maximum likelihood estimates subject to the constraint that $\beta_1 < \beta_2 < \ldots < \beta_G$, and then use these estimates as starting values when maximizing the likelihood without this constraint. This method worked well in simulations.

3 Illustration: Ethnic gap in English test scores at age 16

In this section we apply our methods to re-examine the issue of ethnic gaps in the General Certificate of Secondary Education (GCSE) tests taken by all pupils in England at age 16. The findings in this section should be taken as illustrative, as the analysis is only intended to be a first re-examination of the topic in the light of the new methods introduced in section 2 rather than a comprehensive treatment of the topic.

Previous work suggests that these ethnic gaps are large at age 11 in favour of the White British Majority — with the exception of Chinese students who always do well — and then decrease between age 11 and age 15. By age 16 when children take their GCSEs, ethnic minority pupils have improved to the point of outperforming the White British majority. Only pupils of Black Caribbean descent are reported to still lag behind the White majority at age 16 (Wilson et al. 2005, Connolly 2006, Rothon 2007, Stevens 2007, Strand 2008). These *ethnic group gaps* are policy relevant in the UK due to a public concern that children of immigrants, particularly those of Caribbean descent, have lower academic performance and state schools do not do enough to address their needs (HMSO 1974; 1985).[2]

The two most relevant previous pieces of work on the topic are those of Wilson et al. (2005) and Strand (2008).[3] Wilson et al. use the National Pupil Databese (NPD), which is the set of administrative records for the

[2]From the point of view of the authors no *causal relationship* between ethnic group and school performance can be claimed. As a consequence, we limit ourselves to *describing* the ethnic gaps.

[3]For the sake of brevity we do intend to do here an exhaustive review on the literature on the educational inequalities in the English Secondary Education system. We prefer instead discussing with some detail the pieces of previous work that are the most related to our own and that, we believe, will help the reader to put in context our contribution. For an excellent review of the literature on ethnicity inequalities in the English secondary education system see Stevens (2007).

whole population of students in all state maintained schools in England from 2002 onwards. Strand, on the other hand, uses the Longitudinal Study of Young People of England (LSYPE). The LSYPE is a longitudinal survey of a random sample of pupils in English schools who were in grade 9 (aged 14) in 2004. The NPD is *long but narrow* in the sense that one observes the whole population but only a limited set of variables. In particular, no information on mother's education is available. In contrast, the LSYPE is *short but wide* in the sense that only a random sample is available but a large set of characteristics of the sample individuals and their families is collected. LSYPE contains detailed information on mother's education but is subject to some survey and item non-response.

Wilson et al. (2005) estimates the ethnic gaps in test scores for the whole population of pupils in state maintained schools who took GCSE exams in 2002. Due to the relative scarcity of control variables in the data, Wilson et al. use a dummy indicator of eligibility for Free School Meals (FSM) as the only proxy for family background. The authors find evidence to suggest that Black Caribbean pupils score nearly 0.09 standard deviations below the White British majority. All other minority groups do better than the White British group. In particular, students of Chinese and Indian descent score 0.59 and 0.29 standard deviations above the white reference group. This approach has the advantage of working with the whole population so that the researcher is able to estimate the ethnic group gaps with high precision. However, estimators are likely to be subject to omitted variable bias, or *confounding*, due to the omission of relevant explanatory variables. In other words, the Wilson et al. approach has the disadvantage of potentially *confounding* heterogeneity in background characteristics, such as the mother's education, with systematic unexplained heterogeneity across different ethnic groups — which is the key policy concern.

Strand (2008) estimates ethnic gaps in test scores using the LSYPE and controlling for a large set of family background characteristics, including mother's education, but ignoring potential problems of survey and item non-

response. Among his most important findings the author reports no evidence of significant test score differences between low socio-economic status (SES) boys with White British, Black Caribbean, and Mixed heritage. These boys, nonetheless, do significantly worse than boys from other ethnic groups. Moreover, pupils of Black Caribbean descent from households with medium and high socio-economic status are found to do substantially worse than pupils from other ethnic groups and comparable socio-economic status.

A disadvantage of Strand's approach is that, though the LSYPE oversampled ethnic minorities, the ethnic gaps are necessarily estimated with less precision than those obtained on the basis of NPD because the information available comes only from a random sample of substantially fewer students. Further, and more substantially, Strand's strategy may provide biased estimates if LSYPE unit and item non-response cannot be accounted for by observable differences in individual characteristics *and* if the unobservable heterogeneity is correlated with the GCSE scores. Such *informative selection* may well be a problem in the LSYPE. For instance, mothers of high achieving pupils may participate in the LSYPE at higher rates than mothers of low achieving pupils, because the former are more interested in their children's education.

In section 3.4 we exploit the ability to link NPD and LSYPE to estimate the model described in section 2.

3.1 Data

3.1.1 National Pupil Database

The English National Pupil Database (NPD) is the set of administrative records for the whole population of students in state schools in England.[4]

[4]In the UK independent (non maintained) schools are schools that do not receive public funding and constitute, therefore, the private sector. The private sector is relatively small but important in many respects. Pupils attending non maintained schools represent nearly 7% of the population of all students in England. The NPD also includes some limited

Data from the NPD is available from 2001/2002 onwards and has two components. First, there is information on attainment in the National Curriculum Tests (also known as Key Stage examinations) taken by all children in state schools in England at ages 7, 11, 14 and 16. Second, for maintained schools (state-funded), the Pupil Annual School Census (PLASC) provides some student background variables such as ethnicity and eligibility for free school meals. However, no data on family background such as social class, family income, or parental education are available.

Examinations at age 16 are known as the General Certificate of Secondary Education (GCSE) tests and are graded by independent examination boards regulated by the Qualifications and Curriculum Authority.[5] Mathematics, English, and Science are core subjects and most students attempt these subjects. However, students can take as many GCSEs as they wish.

Our response variable is a test score, known as the capped 'new style GCSE score' (hereafter, GCSE score), which is the sum of scores of up to 8 exams, each score depending on both the difficulty of the exam tier taken and the grade obtained.[6] To account for the fact that students take different numbers of exams, the GCSE score accumulates points obtained on the eight (or equivalent) subjects where the student has performed the best. Points obtained in Mathematics, English, and Science are, therefore, not necessarily included. However, the GCSE score is considered to take a better account of the complex supply of qualifications and options taken by students at age 16 than, for instance, points obtained only in English, Science, and Maths. For this reason GCSE scores are used by the education authorities for Contextualised Value Added (CVA) modeling to measure school performance,

information of pupils in the private sector but here we do not use such information for the analysis.

[5]Besides the exam, which is sent to an independent marker, some GCSE subjects have a *course work* component that is graded by the teacher.

[6]Pass grades are, from top to bottom: A⋆, A, B, C, D, E, F and G. In many subjects, two different "tiers" of examination are offered. A level 1 qualification awards grades D–G whereas a level 2 qualification awards grades A⋆–C.

construct school league tables, and design public policy (see, for example, Ray 2006).

We use data for all year 11 pupils enrolled in maintained English schools in 2006. The cohort reached age 16 that year and, consequently, sat GCSE exams. There were 649,818 year 11 registered students in maintained schools in 2006. Students with a full statement of special education needs (SEN) are excluded from the analytic sample. These are students with severe disabilities and/or learning difficulties and represent around 3.56% of the total population of students in England. After excluding the 23,122 students with full SEN statement, there is a total of 626,630 pupils.[7] Another 62,821 records, around 10% of the population, do not report postcode or any other geographical information. Given that we link the data with neighbourhood (super output area) information from the Census 2001 (see below), these records are also excluded from the analysis. The analytic sample then contains 563,809 cases.

The GCSE score ranges from 0 to 540, with mean 298.42 and standard deviation 101.80.[8] Zeroes are assigned to pupils who were registered for the 2006 academic year and who either did not achieve a pass mark in any of the exams or were reported absent in all examinations. Nearly 2% of the sample have a zero score.

[Table 1 around here]

From PLASC, at pupil level, there is information on date of birth (year and month), gender, ethnicity, eligibility for free school meals, first language spoken at home, special education needs, and a few other variables. Ethnicity is defined in terms of 14 groups and the main source of the information are parents (72.2%), pupils (13%), and school (11.4%) — see Table 2. PLASC

[7] Excluding sixty two females registered in boys schools (single sex) and four males registered in girls schools (single sex). These are clearly misclassified records.

[8] Six missing scores were filled using the GCSE score reported in previous years. These are records of high achievers with mean score 375.16 and who took GCSEs early.

reports a number of school characteristics including school gender mix (girls, boys, mixed), school type,[9] pupil-teacher ratio, percentage of pupils eligible for free school meals, percentage of pupils from the main ethnic groups, and the percentage of pupils whose first language is believed to be English.

[Table 2 around here]

3.1.2 Longitudinal Survey of Young People in England

The Longitudinal Study of Young People in England (LSYPE) is a survey of students, born between 1 September 1989 and 31 August 1990, attending schools in Year 9 in England on February 2004. This cohort reached age 16 in 2006. Here we consider the subset of students who attended maintained schools, thus excluding from the analytic sample students from independent schools and pupil referral units.

The sample design is described in Ward and D'Souza (2008) and briefly summarized here. Schools were the primary sampling units. Maintained schools were stratified into deprived (top quintile in percentage of students qualifying for a free school meal) and non-deprived schools, and deprived schools were oversampled by a factor of 1.5. In the second stage, students were sampled within schools, oversampling major minority groups recorded in PLASC (Indian, Pakistani, Bangladeshi, Black African, Black Carribean and mixed) in order to achieve target issued sample numbers of 1000 in each ethnic group. The school sampling stage took into account the number of students from each of these minority groups. Taken together, the school and student selection probabilities ensured that within a deprivation stratum, each ethnic group had an equal probability of selection (Ward and D'Souza 2008).

Of the 838 maintained schools sampled, 646 (79%) co-operated with the study. Interviewers visited schools to collect the address details of pupils

[9] Main types of State maintained education establishments in England are Community, Voluntary aided, Voluntary controlled, and Foundation schools. See DCSF (2008) for further information.

sampled into the study. They then sent advance letters to both the parent/guardians and young people. The interviews lasted about 1 1/2 hours and consisted of five modules: (1) interview with young person, (2) adult interview for household information, (3) main parent interview, (4) second parent interview and (5) child history. Every young person completing an interview was given a £5 gift voucher.

A total of 20,459 students were sampled into the LSYPE from maintained schools. There were some refusals or unproductive interviews. In the case of refusals, the fact that the student was approached and refused to participate is recorded. For all other unproductive interviews the researcher knows that the student was sampled into the LSYPE but that the interview was unproductive for some reason — 14 specific reasons are known, including broken appointment, no contact, and address inaccessible. For all other records we know when there is a full or a partial interview.

Out of 20,459 sampled students, there are 650 students with full SEN statement (3.17%), 2 females registered in boys schools, 1,039 observations (5.07%) with no postcode or any other geographical information, and 89 observations where the data for the whole household was lost. Excluding these observations we obtain a sample of 18,768 students. From the sample total, there are 12,879 observations (68.63%) with valid response on mother's education (in 493 cases, the mother was not a member of the household, see Table 1).[10] Although we analyze the GCSE results at age 16, corresponding to wave 3 of the LSYPE, we use only data from wave 1 (mother's education), so that attrition is not a problem here.

School non-response was found by Ward and D'Souza (2008) to be related to deprivation status, geographical area (London versus other) and the interaction between deprivation and geographical area. Student non-response was found by Ward and D'Souza to be related to geographic region, ethnicity, qualifications (level achieved in GCSEs) and the interaction between

[10]The LSYPE defines a mother as the natural, step, adoptive or foster mother of the child.

geographic area and being white.

3.1.3 UK 2001 Census

Neighbourhood characteristics at the lower layer super output area (LSOA) level are available from the 2001 UK Census.[11] These data can be linked to NPD records using pupils' postcode information collected in PLASC every year. There are two problems with the linking. First, we used 2006 postcodes from NPD and the last UK Census was carried out in 2001. Hence, when performing the linkage, one needs to assume that the characteristics of the 2001 postcode are a good proxy for the characteristics of the 2006 postcode. Second, the postcodes collected in PLASC are updated regularly. However, some schools may be better than others at keeping the records up to date. Hence, the linkage of NPD and neighbourhood characteristics is far from perfect.

From the Census there is neighbourhood information (% of people in each cell) about social class (professional/skilled manual/unskilled), qualifications (no qualifications/other qualifications/GCSEs/GCE/Higher education no degree/First Degree), unemployment, ethnic group (14 categories), country of birth (7 categories), index of multiple deprivation, and population density.

3.2 Identification strategy

For better identification, we will impose some exclusion restrictions. Specifically, we identified a variable, company that did the LSYPE field work, that enters the selection model only and another variable, an indicator for the child being winter born, that enters the GCSE model only.

[11] The UK Census geography is divided in output area (OA), lower layer super output area (LSOA), middle layer (MSOA), and Government Office Region (GOR). An output area has a minimum size of 100 residents and 40 households. Lower Layer SOAs are aggregations of OAs and have a minimum size 1000 residents and 400 households.

The LSYPE field work was commissioned to the British Market Research Bureau (BMRB) as lead contractor, in association with GFK-NOP and Ipsos-Mori.[12] All three institutions are private companies specialised in market research. Data was collected using computer assisted face to face interviews. Validation of the data collected and enhancement of the study has been undertaken by the National Centre for Social Research (NatCen).[13]

For each student, we know which company attempted / performed the interview. There are four groups: (a) British Market Research Bureau; (b) Ipsos MORI; (c) GfK NOP; (d) joint work BMRB-Mori or NOP-Mori. This is an important piece of information because companies may differ in their experience, ability, effort and incentives to track down and interview individuals. This is particularly relevant because the BMRB was the lead contractor and had more incentive to perform better. Clearly, the company that did the wave 1 fieldwork is likely to affect the probability of selection — i.e., the probability of observing a non missing mother's education entry. However, there are no reasons to believe that the company that did the fieldwork can have an effect on the children's GCSE test score or on mother's education. As a consequence the company that did the LSYPE fieldwork enters the model for S_i but not models for x_i and y_i. Table 3 provides evidence that the company that did the wave 1 field work is a strong predictor of selection.

[Table 3 around here]

According to English law, children must have started school by the beginning of the term following their fifth birthday. However, no minimum age is specified and local education authorities (LEAs) are free to determine the admissions policy for all maintained schools in their area. Moreover,

[12]For further information visit the corresponding web pages: BMRB (http://www.bmrb.co.uk), GFK-NOP (http://www.gfknop.com/customresearch-uk/), and Ipsos-Mori (http://www.ipsos-mori.com/).

[13]NatCen is also a private market research company. Further details at: http://www.natcen.ac.uk/

while there is some room for parents to exercise their discretion, 'anecdotal evidence suggest that there is universal compliance with local admission policies' (Crawford et al. 2007, p. 2). The academic year runs, in all cases, from 1 September to 31 August and children are rarely held back even if they do not reach key academic targets.

These institutional arrangements mean that 'there is considerable geographic variation in the age at which children born on a particular day of the year start primary school' (Crawford et al. 2007, p. 2). The age differential is particularly exacerbated for children who are born just before and after the summer holidays. Children who were born on August 31st, for instance, may enter school in September when they are 4, or in September when they are 5. In contrast, children who were born on September 1st enter school in September when they are 5. In other words, due to local education authority policy, a child born in the summer may end up entering school almost a year ealier than the eldest pupil in her/his cohort.[14]

The existence of age differences in primary school entry in England is a well known fact and various studies have investigated the relationship between age and academic performance. Thomas (1995) looks at 1991 Key Stage 1 tests and finds evidence that older children in a year group perform better at Key Stage 1 exams than younger pupils. Similar results are reported by Bell and Daniels (1990) in a study of performance in a science test for English children aged 11, 13, and 15. Finally, Crawford et al. (2007) use NPD data for various cohorts of children and show that age within cohort has a positive and significant effect on school performance at ages 16 and 18. The effect of age on school attainment has also been investigated in countries other than England. In Germany, for instance, Puhani and Weber (2007) find that children who entered school aged 7 perform better in a literacy test taken at the end of primary school than children who entered school aged 6.

From the NPD we have knowledge of the exact date of birth (year and

[14] Crawford et al. (2007) give further details on how admissions policy vary across different local education authorities.

month) of each student in England. We therefore define a dummy variable to identify pupils born during the autumn and winter months (September 1st to December 31st). These pupils generally enter primary school before the age of 5. Further, this winter born dummy variable is likely to affect a child's GCSE score but we claim it does not affect either mother's education or selection, and thus can be used to specify an exclusion restriction. In particular, the winter born dummy variable enters the model for the test score y_i but not models for x_i or S_i. Table 4 shows that the winter born dummy variable is a good predictor of the GCSE score. In fact, children born in the winter months achieve a GCSE score almost six points higher than non winter born children.

[Table 4 around here]

There are two main threats to our identification strategy with the winter born dummy variable. First, there is the possibility that women who are more willing to participate in the LSYPE and answer the education question may have children in the winter months at higher rates than women who refuse to participate. Second, highly educated mothers may tend to have children in the winter months at higher rates than less educated mothers. However, simple tabulations of the data in Tables 4 and 5 show that women with winter born children are not significantly different to women with no winter born children as far as education and selection refers. As a consequence, the exclusion restriction seems reasonable.

[Table 5 around here]

3.3 Descriptive results

Table 1 reports some descriptive statistics for the capped GCSE new style point score (main response, y_i) and the selection variable (S_i). There are 18,679 individuals who were sampled into the LSYPE, representing 3.3% of the cohort. The education of the mother is observed for nearly 72% of the

pupils sampled in the LSYPE. In other words, there is a non response rate of about 28%. As we discussed earlier in the text, the mother's education is sometimes missing due to item non-reponse and sometimes due to unit non response. There are some differences in the mean GCSE scores across the three groups. In particular, pupils who were sampled into the LSYPE and have $S_i = 1$ perform around 10 points higher than those sampled into the LSYPE for which the education of the mother is missing. Hence, simply by examining the raw data the reader may conclude that selection is likely to be informative. Details on the distribution of education among mothers for whom the key control x_i is observed are presented in Table 6.

[Table 6 around here]

Descriptive statistics by ethnic group are presented in Table 7. There are 14 ethnic categories, including one category for those who actively refused to state their ethnic background and one category for those cases where there is no ethnic information available. Most pupils belong to the White British group. As a whole, this group represents 82% of the population (see column 1). There are relatively large differences in mean GCSE test score. On average, pupils with Chinese ethnic background are the best performers with an average test score of 361 points. At the opposite end, pupils with Black Caribbean background score on average 272 points. That is, there is a difference of 89 points between the group highest and lowest mean score (see column 3). Pupils sampled into the LSYPE with White British background represent nearly 2.05% of the total population of students in England (column 4 dived by 10). Students with Indian, Pakistani, and Mixed ethnicity are the largest minority groups and none of them represent more than 0.23% of the total population. Column 5 of Table 7 presents the response rate for each ethnic group. This response rate is, effectively, the proportion of pupils sampled into the LSYPE for whom we observe their mother's education. From this column the reader can easily conclude that there is a wide variation in the response rate across the different ethnic groups, ranging from 51% for

the Chinese group to 74% for the White British majority group.

[Table 7 around here]

The last two columns in Table 7 reports the proportion of pupils whose mother completed GCSE qualifications or higher. One column gives unweighted statistics whereas the other gives weighted statistics using the LSYPE probability weights. Pupils with Bangladeshi background stand out as the most disadvantaged as far as the mother's education is concerned. In fact, for this ethnic group raw data indicates that only 11% (weigthed) of students belong to a family where the mother has at least some GCSE qualifications. Pakistani and Chinese pupils are also clearly disadvantaged, though not as much as those of Bangladeshi origin. Another important observation is the relatively high proportion of mothers of students of Black Caribbean descent who report having at least GCSE qualifications; nearly 82% (weigthed). This figure is the highest among all the 14 groups and is only comparable with the one reported for the White British majority. The White British majority, however, lags behind the Black Caribbean group by about 7.53 percentage points.

[Tables 8 and 9 around here]

3.4 Ethnic gaps estimates

We move now to discuss the main results, which are presented in Tables 8 and 9. As previously discussed in section 2, we fit a system of three equations to the data: one equation for the selection dummy variable (S_i), one equation for the main response variable (y_i), and one equation for the missing ordinal covariate (x_i). The covariates included in all three equations are listed and described in Table 10. Table 8 reports coefficients in the main response equation (i.e., the GCSE score). To ease interpretation, the GCSE score was standardised so that coefficients can be interpreted as changes in terms of standard deviations. Details of the coefficients estimated for the S_i and

x_i equations in the missing covariate model are reported in Table 9. The missing covariate model was estimated by Maximum Simulated Likelihood with 800 Halton draws. Adding 100 more draws did not cause important changes on coefficients and/or standard errors.

For comparison purposes Table 8 also reports results from linear regressions of the standardised GCSE score on covariates, fitted to NPD data alone using ordinary least squares (OLS) and fitted to LSYPE data alone using weighted least squares (WLS). Obviously, the OLS regression fitted to NPD alone cannot control for mother's education and the estimators of the ethnic gap are subject to simple confounding. In contrast, the WLS regression fitted to LSYPE controls for mother's education but uses exclusively the *selected sample*. In other words, cases for which mother's education is missing are listwise deleted so that the WLS estimators are subject to potential selection bias, as well as having larger standard errors.

Two different specifications of the missing covariate model were fitted to the data. A *benchmark* model/specification uses only information that is available from NPD and LSYPE, using the identification strategy discussed in subsection 3.2. Next, a set of extra control variables from the 2001 UK Census, measured at the lower layer super output area level (which are small local geographic areas, see footnote 11), were added to the benchmark specification and results reported in the last two columns of Table 8 and in the right hand side panel of Table 9. The additional controls include: population density, index of multiple deprivation, % of population in the LSOA with a given level of education (6 groups, same groups as those in Table 6), and country of birth (7 groups). The same identification strategy that we use in the benchmark specification is used in the *extra controls* specification. Note, however, that the authors are aware that some or all of these extra variables are potentially correlated with ϵ_{yi} and, as a consequence, the researcher should take the findings from this latter specification with caution. The extra control variables are likely to be good predictors of all three: (i) children's achievement, (ii) mother's education, and (iii) probability of selection (i.e.,

probability of observing mother's education). As a consequence, we let these variables enter all three equations in the missing covariate model. Hence, the value of fitting the missing covariate model with these extra control variables is simply investigating how findings would change if a set of extra predictors (some potentially endogenous) are exploited. As a consequence, the reader should take the results from the *extra controls* specification as a robustness check of the benchmark specification. Table 10 gives detail of the variables that enter all equations in the benchmark and extra controls specifications.

[Table 10 around here]

Before discussing the ethnic gaps in GCSE tests scores it is important to mention that the Winter born dummy variable is found to be a strong predictor of performance — see bottom part of Table 8. In fact, a test for the exclusion of the Winter born dummy variable in the GCSE score equation is easily rejected with $\chi^2(1) = 954.36$ (p-value < 0.001). Similarly, a test for the exclusion of the three LSYPE-interviewer company dummy variables in the selection equation rejects the null with a $\chi^2(3) = 64.99$(p-value < 0.001) (see Table 9). That is, the LSYPE-interviewer company dummy variables are strong predictors for selection. Hence, if the relevant orthogonality conditions hold, there is little reason to suspect that the missing covariate model may suffer from a problem of tenuous identification (for more on this, see Keane 1992).

Looking at Table 8 the reader can see that, as expected, we find evidence of a negative $\text{Cor}(\epsilon_{xi}, \epsilon_{Si})$ and a positive $\text{Cor}(\epsilon_{yi}, \epsilon_{Si})$. Walds test for $\text{Cor}(\epsilon_{xi}, \epsilon_{Si})$ and $\text{Cor}(\epsilon_{yi}, \epsilon_{Si})$ based on the *arctanh* transformation give $z = 16.82$ [p-val < 0.001] and $z = -15.09$ [p-val < 0.001], respectively. Hence, the two correlations are different from zero at the 1% level of significance. The result of a negative $\text{Cor}(\epsilon_{xi}, \epsilon_{Si})$ implies that, after controlling for $\mathbf{z}$, mother's education is less likely to be observed when it takes larger values. Similarly, $\text{Cor}(\epsilon_{yi}, \epsilon_{Si}) > 0$ is consistent with the hypothesis that a mother's education is observed with higher probability when her child is a

high achiever. We use this basic intuition in the discussion of the GCSE ethnic gaps, which we now introduce. In what follows the reader should keep in mind that, comparing the size of the standard errors across the models presented in Table 8, there are large efficiency gains to be obtained by using the whole NPD instead of just using the LSYPE.

Except for the Black Caribbean and Black Other groups, all minorities do better than the White British reference group in their GCSE scores. This observation is true regardless of the strategy used to estimate the ethnic gaps. There are three ethnic groups for whom conclusions drawn from the missing covariate model differ from those for the linear regressions: (i) Bangladeshi; (ii) Chinese; and (iii) Black Caribbean.

OLS regression fitted to the whole population using data only from NPD suggest that Bangladeshi pupils score on average 0.40 standard deviations (SDs) higher than the White British majority (see first column of Table 8). This figure is very similar to the 0.48 SDs gap reported by Wilson et al. (2005) estimated from the NPD for the cohort of pupils that were 16 in 2002. Obviously, this figure is subject to confounding, or omitted variable bias, due to the fact that a key control, mother's education, has not been accounted for. Moving to the second column of Table 8 the reader can find the gap estimated by WLS fitted exclusively to the subset of pupils sampled into LSYPE for whom we observe the education of the mother. Now, the Bangladeshi/White British gap is found to be of the order of 0.69 SDs. We have no comparable gap from Strand (2008) due to the fact that the author uses total points score rather than capped new style score. However, Strand reports a gap of 0.56 SDs using the total points score and a much richer set of controls (see Strand, Table 3.2). Hence, the two figures broadly agree. Obviously, these estimates are naive in the sense that WLS regressions are fitted on the *selected* LSYPE sample and are potentially biased because mother's education is missing according to an informative selection rule.

The missing covariate model introduced in section 2 exploits the linkage of NPD, LSYPE, and 2001 UK Census, to control for mother's edu-

cation while addressing selection on unobservables that may be correlated with achievement and mother's education. Column 3 of Table 8 presents coefficients obtained from the missing covariate model and the benchmark specification (using only linkage between NPD and LSYPE). Now, the gap between Bangladeshi and the White British majority is estimated to be 1.61 SDs in favour of the former ethnic group. Hence, the difference between the gap estimated by OLS fitted to NPD alone and the gap estimated using the missing covariate model is larger than 1 SD. Clearly, this is a sizeable correction. This correction may seem too large. However, Table 7 shows that Bangladeshi pupils are considerably less likely to have mothers with GCSE or higher qualifications. Despite this disadvantage, children from the Bangladeshi group attain a mean GCSE score that is less than 0.5 points lower than the White British majority. As a consequence, one would expect a large positive correction due to controlling for mother's education. Besides, the correction not only accounts for simple confounding but also for controlling for potential informative selection.

The case of students of Chinese descent goes in the opposite direction. OLS regression fitted on NPD reports a gap of 0.60 SDs between Chinese children and the White British control group. Now, fitting the model using WLS and the *selected* sample of LSYPE would suggest that the gap is even higher, at about 0.86 SDs. However, the missing covariate model suggest a much more modest positive gap of 0.47 SDs. In other words, the missing covariate model suggests in this case a negative correction of about 0.13 SDs. We turn again to Table 7 to offer some intuition to explain this result. Descriptive statistics show that Chinese pupils have less educated mothers than the White British majority (though not as disadvantaged as the Bangladeshi). This would suggest a positive correction. However, among those sampled into the LSYPE, the proportion of Chinese pupils for whom we know their mother's education is by far the lowest among all ethnic groups (only 51%, see column 5 of table 7). If the probability of observing mother's education is negatively correlated with the value that her education takes when miss-

ing, one would expect then that the education of the mother of a Chinese pupil will be seriously underestimated. As a consequence, once informative selection is accounted for, it is intuitive to find that a negative correction in the Chinese/White British score gap is needed because $\text{Cor}(\epsilon_{xi}, \epsilon_{Si}) < 0$ is empirically reported in the bottom panel of Table 8. Also, the negative correction in the Chinese / White British gap may be due to the low response rate among Chinese mothers because $\text{Cor}(\epsilon_{yi}, \epsilon_{Si}) > 0$ makes it appear as if poor performers were under represented among Chinese students. This is an example of how, compared to linear regression strategies, the missing covariate model allows the researcher to gain a better understanding of what the explained and unexplained ethnic group gaps are in GCSE scores.

Finally, the missing covariate model suggests that the White British / Black Caribbean gap is slightly underestimated by the OLS regression based only on NPD and slightly overestimated by the WLS regression based only on the *selected* sample of LSYPE. The correction is, nonetheless, rather small. Again, the result is intuitive. Firstly, the reader can see from Table 7 that the proportion of mothers with GSCE or higher qualifications is very similar among pupils with Black Caribbean background and the White British majority. Black Caribbean mothers are, if anything, slightly better educated. Hence, by simple confounding (or excluded variable bias), one would expect that after controlling for mother's education the test score gap will be corrected downwards. However, the correction is not expected to be too large because the differences in mother's education among Black Caribbean mothers and White British mothers are relatively small; only 7.5 percentage points. This leaves, in turn, little room to allow for any correction due to the informative selection rule.[15] In other words, intuition suggests that estimates of the White British / Black Caribbean gap based on the missing covariate model and the WLS regression using the *selected* LSYPE sample should not be too different.

[15] The $\text{Cor}(\epsilon_{xi}, \epsilon_{Si}) < 0$ would tend to re-inforce the negative correction that is induced by simple confounding but the $\text{Cor}(\epsilon_{yi}, \epsilon_{Si}) > 0$ will play in the opposite direction.

Column 4 of Table 8 reports results from the missing covariate model and the *extra controls* specification. This specification exploits linkage of three datasets, NPD, LSYPE, and 2001 UK Census. As we discussed earlier, the same identification strategy as before is used here. Hence, the only difference between the benchmark specification and the extra control specification is that some extra control variables from the 2001 UK Census are added to all equations. Clearly, inspection of column 4 of Table 8 shows that adding the extra controls changes coefficients only marginally in most cases. Probably the only case that deserves mentioning is the Bangladeshi/White British difference, where the estimated ethnic gap goes from 1.6 SDs in the benchmark specification to just 0.69SD in the specification with extra controls. Hence, adding the extra predictors allows us to halve the gap. Note, however, that the Bangladeshi/White British ethnic gap in GCSE scores is still the largest positive gap recorded, and a positive correction is suggested by the missing covariate model with respect to the gaps reported by OLS fitted on NPD-only data and WLS fiited on the *selected* LSYPE sample. Hence, conclusions remain the same.

A final point that is worth mentioning is the fact that the coefficient on the Free School Meals dummy variable goes down once the mother's education is controlled for and informative selection is accounted for (see coefficient on the dummy variable across the four models in Table 8). This is, obviously, what one expects to see as some of the effect that is wrongly attributed to income in the linear regressions will be rightly labelled as variation due to family background heterogeneity (selection bias) in the missing covariate model.

4 Discussion

This paper considers the problem of parameter estimation in a model for a continuous response variable when an important ordinal explanatory variable is missing for a large proportion of the sample and selection into missingness is informative — i.e., data are not missing at random (NMAR). We suggest

solving the endogenous selection problem by modelling the selection mechanism, the ordinal explanatory variable, and the response variable together.

We use our methods to re-examine the 'problem' of describing the *ethnic group gaps* in the General Certificate of Secondary Education (GCSE) test scores at age 16 in England using the National Pupil database (NPD). The NPD contains administrative test results for the entire population of pupils in state schools in England. However, only a limited set of controls are available and there is no information on mother's education. Mother's education is an important background characteristic that researchers and policy makers would like to control for to estimate *unexplained ethnic gaps* in test scores at age 16. We exploit the ability of linking individual records in the NPD with individual records in the Longitudinal Study of Young People in England (LSYPE), which is a survey of year 9 students attending schools in England in 2004. The LSYPE contains detailed background information, including mother's education. The linkage is valuable because it allows us to know the value of mother's education for a subset of the pupils in the NPD. The LSYPE, however, is subject to some unit and item non-response (around 28%). Further, we suspected that mothers of high achieving children will be willing to respond LSYPE's questions at a higher rate than mothers of low achieving children. Similarly, we suspected that busy professional mothers will be harder to track and less willing to answer LSYPE's questions. As a consequence, there were good reasons to believe that selection into missingness is *informative*. Our method allows us to estimate the ethnic group gaps, net of differences in background characteristics including mother's education, while exploiting GCSE and ethnicity data on the entire population.

Our study has a number of limitations. Importantly, we suppose that a set of good predictors of mother's education are available for the whole population. This may be challenging as administrative data typically contain a limited number of background characteristics, which is why mother's education is missing in the first place. We used child's ethnicity, eligibility to free school meals (a proxy for socio-economic status), type of school, and

geographic dummies as predictors of mother's education in our benchmark specification. Then, we complemented this list with neighbourhood characteristics at the lower layer super output area (LSOA) level from the 2001 UK Census. There is no need to suppose that there is a causal relationship between these predictors and mother's education, as the equation for mother's education in our model does not need to have an structural interpretation. The only requisite is that these predictors are useful for predicting mother's education and are uncorrelated with ϵ_{yi} . Obviously, the higher the correlation between these predictors and mother's education the better.

Another important issue is the possibility that the missing covariate model key variables, mother's education and child's ethnicity may be subject to some misclassification. In the case of ethnicity, tabulations of the NPD ethnicity data across various years for the cohort of children who were the target population of the LSYPE show that most of the ethnicity entries remain consistent / unchanged over time. Moreover, the LSYPE offers an independent source of ethnicity for the children who were sampled into the survey. For these children, tabulations show that the ethnicity entry in NPD and LSYPE are highly consistent. Hence, the evidence suggests that there are no reasons to believe that the NPD ethnicity data are subject to serious misclassification. In the case of mother's education we were unable to perform a similar check as the variable is only recorded once in the LSYPE.

Our standard errors were based on the outer product of gradients and did not take the nesting of students in schools or neighborhoods into account. Calculation of Eicker–Huber–White robust standard errors for clustered data is theoretically possible. However, this requires obtaining the Hessian matrix either analytically or numerically. Analytical second derivatives are in practice unavailable given the complexity of the likelihood function. Numerical calculation of the Hessian is feasible for data sets that are relatively small, i.e., in the order of a few thousand observations. In our application, however, the sample size is nearly six hundred thousand individuals and maximisation involves the handling of around 1.6 million equation-person data points. In

cases like this, numerical evaluation of the Hessian matrix is unfeasible and the researcher must rely exclusively on the OPG approximation.

Finally, it is important to say that our model deals with the problem of a single missing ordinal covariate. Obviously, there are other important missing background characteristics, besides mother's education, that we did not control for. These include father's education and family income. Dealing with two or more missing covariates at the same time demands writing a model that, given its complexity, will be unfeasible to estimate in practice.

Despite the aforementioned limitations, we believe that the model presented here is a valuable methodological device to address the problem of a missing ordinal covariate with informative selection when the researcher has the ability of linking administrative and survey data (or, in general, two or more complementary datasets). In our application we offer an example where sample selection is informative. Hence, analysis of the survey data alone leads to inconsistent estimators. Furthermore, analysis of the administrative data alone is unsatisfactory because estimators are subject to relevant omitted variable bias due to omission of mother's education, a control variable that is important to control for in order to set apart explained from unexplained differences in test scores at age 16 for different ethnic groups. Our findings from the missing covariate model suggest that for groups like Bangladeshis, failing to control for mother's education may seriously overestimate the extent of the ethnic disadvantage. For other ethnic groups results go the other way. Black Carribbeans, for example, look more disadvantaged once mother's education is controlled for. And the Chinese ethnic advantage that has been reported in previous literature is significantly reduced though it remains positive.

References

Bell, J., Daniels, S., 1990. Are summer-born children disadvantaged? the birthdate effect in education. Oxford Review of Education 16, 67–80.

Berndt, E., Hall, B., Hall, R., Hausman, J., 1974. Estimation and inference in nonlinear structural models. Annals of Social Measurement 3, 653–665.

Carpenter, J. R., Kenward, M. G., Vansteelandt, S., 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. Oxford Review of Education 32, 235–252.

Chen, H. Y., 2004. Nonparametric and semiparametric models for missing covariates in parametric regression. Journal of the American Statistical Association 99, 1176–1189.

Connolly, P., 2006. Summary statistics, educational achievement gaps and the ecological fallacy. Oxford Review of Education 32, 235–252.

Crawford, C., Dearden, L., Meghir, C., 2007. When you are born matters: the impact of date of birth on child cognitive outcomes in England, Center for the Economics of Education dicussion paper No. 93.

DCSF, 2008. The composition of schools in England, Department for children, schools and families. Statistical bulletin No. B02/2008.

Diggle, P. J., Kenward, M. G., 1994. Informative drop-out in longitudinal data analysis (with discussion). Journal of the Royal Statistical Society, Series C 43, 49–93.

Gelman, A., King, G., Liu, C., 1998. Not asked and not answered: Multiple imputation for multiple surveys. Journal of the American Statistical Association 93, 846–857.

Gourieroux, C., Monfort, A., 1993. Simulation-based inference: A survey with special reference to panel data models. Journal of Econometrics 59, 5–33.

Griliches, Z., Hall, B. H., Hausman, J. A., 1978. Missing data and self-selection in large panels. Annales de Linsee No 30-31, 137–176.

Hausman, J. A., Wise, D. A., 1979. Attrition bias in experimental and panel data: The Gary income maintenance experiment. Econometrica 47, 455–473.

Heckman, J. J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

HMSO, 1974. Education disadvantage and the educational needs of immigrants: observations ot the report on education of the select committee on race and immigration, (Cmnd 5720) HMSO, London.

HMSO, 1985. Education for all: the report of the committee of enquiry into the education of children from ethnic minority groups (chairman: Lord swann), (Cmnd 9453) HMSO, London.

Horton, N. J., Laird, N. M., 1998. Maximum likelihood analysis of generalized linear models with missing covariates. Statistical Methods in Medical Research 8, 37–50.

Huang, L., Chen, M.-H., Ibrahim, J. G., 2005. Bayesian analysis of generalized linear mixed models with nonignorably missing covariates. Biometrics 61, 767–780.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., 2002. Bayesian methods for generalized linear models with covariates missing at random. Canadian Journal of Statistics 30, 55–78.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., Herring, A. H., 2005. Missing-data methods for generalized linear models: A comparative review. Journal of the American Statistical Association 100, 332–346.

Jackson, C. H., Best, N. G., Richardson, S., 2009. Bayesian graphical models for regression on multiple datasets with different variables. Biostatistics 10, 335–351.

Keane, M. P., 1992. A note on identification in the multinomial probit model. Journal of Business & Economic Statistics 10(2), 193–200.

Lipsitz, S. R., Ibrahim, J. G., Chen, M.-H., H. Peterson, H., 1999. Non-ignorable missing covariates in generalized linear models. Statistics in Medicine 18, 2435–2448.

Little, R., Schluchter, M., 1985. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 72, 497–512.

Little, R. J. A., 1988. Missing-data adjustments in large surveys. Journal of Business and Economic Statistics 6, 287–296.

Little, R. J. A., 1992. Regression with missing X's: A review. Journal of the American Statistical Association 87, 1227–1237.

Little, R. J. A., Rubin, D. B., 2002. Statistical Anaysis with Missing Data, Wiley, Hoboken, NJ.

Puhani, P., Weber, A., 2007. Does the early bird catch the worm? instrumental variable estimates of educational effects of age of school entry in Germany. Empirical Economics 32, 359–386.

Ray, A., 2006. School value added measures in england, Tech. rep., Department for Education and Skills.

Robins, J. M., Hsieh, F., Newey, W., 1995. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. Journal of the Royal Statistical Society, Series B 57, 409–424.

Rothon, C., 2007. Can achievement differentials be explained by social class alone? Ethnicities 7, 306–322.

Rotnitzky, A., Robins, J. M., 1995. Semiparametric regression estimation in the presence of dependent censoring. Biometrika 82, 805–820.

Rubin, D. B., 2002. Multiple Imputation for Nonresponse in Surveys, Wiley, New York.

Schafer, J. L., 2002. Analysis of Incomplete Multivariate Data, Chapman & Hall, London.

Stevens, P. A. J., 2007. Researching race/ethnicity and educational inequality in english secondary schools: A critical review of the research literature between 1980 and 2005. Review of Educational Research 77, 147–185.

Strand, S., 2008. Minority ethnic pupils in the longitudinal study of young people in england: Extension report on performance in public examinations at age 16, Tech. Rep. DCSF-RR029, Department for children, schools and families.

Stubbendick, A. L., Ibrahim, J. G., 2003. Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. Biometrics 59, 1140–1150.

Takeuchi, K., Yanai, H., Mukherjee, B. N., 1982. The Foundations of Multivariate Analysis, Wiley, New Delhi.

Thomas, S., 1995. Considering primary school effectiveness: an analysis of 1992 key stage 1 results. The Curriculumn Journal 6, 279–295.

Train, K., 2003. Discrete choice methods with simulation, Cambridge university press.

Vach, W., Blettner, M., 1995. Logistic regression with incompletely observed categorical covariates – investigating the sensitivity against violation of the missing at random assumption. Statistics in Medicine 14, 1315–1327.

Ward, K., D'Souza, J., 2008. LSYPE user guide to the datasets: Wave one to wave three, Tech. rep., Department for children, schools and families.

Wilson, D., Burgess, S., Briggs, A., 2005. The dynamics of school attainment of england's ethnic minorities, Centre for Market and Public Organisation Working Paper No. 05/130.

Wooldridge, J., 2007. Inverse probability weighted estimation in general missing data problems. Journal of Econometrics 141, 1281–1301.

Wooldridge, J. M., 2002. Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, MA.

Wu, M. C., Carroll, R. J., 1988. Estimation and comparison of change in the presence of informative right censoring by modeling the censoring process. Biometrics 44, 175–188.

Zhao, Y., 2009. Regression analyiss with covariates missing at random: A piece-wise nonparametric model for missing covariates. Communications in Statistics – Theory and Methods 38, 3736–3744.

Tables and Figures

Table 1. Selection Variable (S_i) and Capped GCSE new style point score (y_i)

Category	Symbol	S_i	$\bar{y}$	Freq.	%NPD	%LSYPE
Not LSYPE sampled	$x_i\overline{\text{o}}$, $S_i\overline{\text{o}}$	missing	298.40	545,130	96.69	0
LSYPE sampled, respondent	$x_i\text{o}$, $S_i=1$	1	302.46 (299.22^w)	13,372†	2.37	71.59
LSYPE sampled, non-respondent	$x_i\overline{\text{o}}$, $S_i=0$	0	290.10 (299.77^w)	5,307	0.94	28.41
Total			563,809			

† For 493 of these cases, x_i is missing although $S_i = 1$ because mother was reported to be *"not a member of the household"* but survey was otherwise completed.

^{w}Indicates that the statistic has been calculated using probability weights for the LSYPE.

Table 2 Source of ethnic group classification

Category	Freq.	%	Cum.
Current school	64,423	11.43	11.43
Other	11,186	1.98	13.41
Parent	407,147	72.21	85.62
Prev. school	7,380	1.31	86.93
Pupil	73,670	13.07	100.00
Blank on Data	3	0.00	100.00
Total	563,809		

Table 3. Company doing LSYPE field work

Category	Freq.	%	%S=1
BMRB	8,061	43.16	73.63
NOP	8,316	44.52	71.90
Mori	2,183	11.69	64.64
BMRB-Mori or NOP-Mori	119	0.64	39.50
Total	18,679		

Table 4. Winter born children

	Winter born		No winter born	
Category	Obs	%	Obs	%
Not LSYPE sampled	368,744	96.69	176,386	96.68
LSYPE sampled, respondent	9,086	2.38	4,286	2.35
LSYPE sampled, non-respondent	3,527	0.92	1,780	0.94
Total	381,357		182,452	
$\bar{y}$	302.41		296.52	

Table 5. Mothers' education and winter born children

	No winter born			Winter born		
Category	Obs	%	$\%^w$	Obs	%	$\%^w$
1. No qualification	2,299	26.28	19.52	1,152	27.89	20.49
2. Other qualifications	839	9.59	10.63	376	9.43	10.21
3. GCSE grades A-C or equiv	2,659	29.30	33.73	1,210	30.04	32.99
4. GCE A level or equiv	1,091	12.47	13.76	495	11.99	13.34
5. Higher education no degree	1,025	11.72	12.35	514	12.45	12.93
6. Degree or equivalent	836	9.56	10.01	383	9.27	10.04
Total	8,749			4,130		

^{w}Indicates that the statistic has been calculated using probability weights for the LSYPE.

Table 6. Mothers' education, ordinal x_i

Category	Freq.	%	$\%^w$	$\bar{y}$	$\bar{y}^w$
1. No qualification	3,451	26.80	19.83	271.28	252.61
2. Other qualifications	1,215	9.43	10.5	278.60	272.24
3. GCSE grades A-C or equiv	3,869	30.04	33.49	302.82	298.69
4. GCE A level or equiv	1,586	12.31	13.63	323.21	321.88
5. Higher education no degree	1,539	11.95	12.54	333.54	333.22
6. Degree or equivalent	1,219	9.47	10.02	366.76	368.10
Total	12,879				

^{w}Indicates that the statistic has been calculated using probability weights for the LSYPE.

Table 7. Ethnic group

Category	Freq.	%	$\bar{y}$	S_i		x_i	
				$10\%S_i$o	$\%\frac{S_i=1}{S_i\text{o}}$	$\%(\geq 3)$	$\%(\geq 3)^w$
White british	461,070	81.78	298.47	20.46	73.65	73.48	73.35
White other	13,168	2.34	306.93	0.53	67.45	53.61	54.89
Mixed	12,596	2.23	294.99	1.91	70.34	67.99	69.31
Indian	13,061	2.32	334.88	2.10	72.76	46.67	47.95
Pakistani	13,083	2.32	288.33	2.14	68.69	20.67	21.14
Bangladeshi	5,516	0.98	297.92	1.65	68.14	10.54	10.77
Other asian	3,909	0.69	317.65	0.20	71.30	50.62	50.55
Caribbean	8,062	1.43	271.64	1.49	62.98	79.76	81.22
African	9,703	1.72	285.22	1.50	63.83	53.36	52.39
Other black	2,481	0.44	272.69	0.13	62.16	70.73	74.04
Chinese	2,028	0.36	361.65	0.09	50.94	32.00	31.59
Any other	4,931	0.87	285.57	0.23	67.44	32.53	28.87
Refused	6,545	1.16	297.44	0.27	68.39	82.18	83.25
No data	7,656	1.36	277.90	0.43	74.79	67.26	67.39
Total	563,809						

^{w}Indicates that the statistic has been calculated using probability weights for the LSYPE.

Table 8 Estimates for the missing covariate model: standardised capped new style GCSE score equation

	linear regressions				Missing covariate model[a,d,e]			
	NPD[b]		LSYPE[a,c]		Benchmark		Extra controls	
Variable	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Ethnic group (White British)								
White other	0.116‡	0.008	0.384‡	0.059	0.159‡	0.006	0.129‡	0.006
Mixed	0.054‡	0.009	0.023	0.040	0.072‡	0.005	0.072‡	0.005
Indian	0.415‡	0.009	0.513‡	0.033	0.388‡	0.005	0.345‡	0.005
Pakistani	0.232‡	0.009	0.468‡	0.041	0.358‡	0.005	0.342‡	0.005
Bangladeshi	0.407‡	0.013	0.686‡	0.051	1.608‡	0.006	0.697‡	0.007
Asian other	0.282‡	0.015	0.326†	0.110	0.270‡	0.010	0.250‡	0.010
Black Caribbean	-0.106‡	0.011	-0.201‡	0.049	-0.183‡	0.007	-0.170‡	0.007
Black African	0.122‡	0.010	0.232‡	0.053	0.114‡	0.006	0.111‡	0.006
Black other	-0.086‡	0.019	-0.187†	0.146	-0.144‡	0.015	-0.125‡	0.014
Chinese	0.606‡	0.021	0.860‡	0.159	0.473‡	0.014	0.467‡	0.013
Any other	0.095‡	0.014	0.464‡	0.104	0.167‡	0.009	0.141‡	0.009
Refused	-0.028†	0.012	-0.113	0.105	-0.009‡	0.009	-0.022‡	0.009
No data	-0.223‡	0.011	-0.056	0.075	-0.100‡	0.008	-0.094‡	0.008
Mother education								
No qualifications			-0.142	0.054	-1.415‡	0.005	-1.369‡	0.018
Other qualifications			0.039	0.054	0.392‡	0.008	0.411‡	0.019
GCSE A-C			0.281‡	0.049	0.517‡	0.006	0.509‡	0.018
GCE A level			0.466‡	0.050	0.570‡	0.008	0.553‡	0.019
Some higher education			0.565‡	0.053	0.589‡	0.007	0.572‡	0.019
Degree			0.870‡	0.051	0.646‡	0.008	0.648‡	0.019
Other controls								
winterborn	0.056‡	0.003	0.056‡	0.019	0.059‡	0.002	0.060‡	0.002
Free School Meals	-0.616‡	0.004	-0.511‡	0.033	-0.290‡	0.003	-0.217‡	0.003
Geographic Region	Yes		Yes		Yes		Yes	
Gender × School type	Yes		Yes		Yes		Yes	
Deprived School	Yes		Yes		Yes		Yes	
2001 Census variables	No		No		No		Yes	
$\text{Cor}(\epsilon_{yi}, \epsilon_{Si})$					0.169‡	0.010	0.175‡	0.010
$\text{Cor}(\epsilon_{xi}, \epsilon_{Si})$					-0.226‡	0.014	-0.301‡	0.014
$\text{Var}(\epsilon_{yi})$					0.444‡	0.001	0.414‡	0.001
N. observations	563,809		12,879		563,809		563,809	

Note: ‡(†) Significant at 1% (5%). OPG standard errors reported. Dependent variable is the standardised capped new style GCSE score. (a) To ease comparison across columns these models have no constant term to ensure that coefficients on mother's education can be interpreted as the mean when other controls are zero. The coefficients on mother's education are also the locations of the discrete latent variable η_{1i}. (b) Ordinary least squares regression (c) Weighted least squares regression. (d) Estimation was performed by Maximum Simulated Likelihood with 800 Halton draws. Adding 100 more draws did not cause important changes on coefficients and/or standard errors. (e) Details on coefficients in selection and missing covariate equations are given in Table 9.

Table 9 Estimates for the missing covariate model: selection and mother's education equations

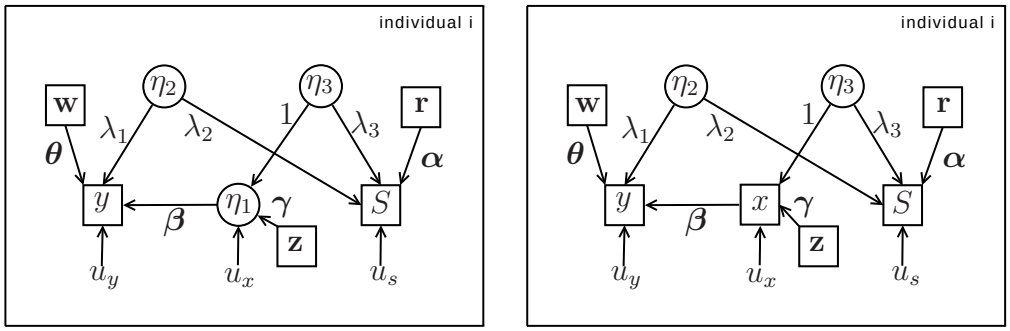
	Benchmark				Extra controls			
	Selection Equation		Missing Covariate (Mother's education)		Selection Equation		Missing Covariate (Mother's education)	
Variable	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
LSYPE survey company (BMRB)								
NOP	-0.069‡	0.021			-0.076‡	0.021		
Mori	-0.173‡	0.034			-0.174‡	0.035		
Other	-0.737‡	0.112			-0.713‡	0.112		
Ethnic group (White British)								
White other	-0.172†	0.080	-0.066‡	0.015	-0.181†	0.080	-0.069‡	0.014
Mixed	-0.024	0.042	-0.020	0.012	-0.011	0.042	0.003	0.013
indian	-0.049	0.045	0.040‡	0.012	-0.092	0.048	0.059‡	0.012
Pakistani	-0.194‡	0.047	-0.202‡	0.012	-0.238‡	0.052	-0.160‡	0.013
Bangladeshi	0.000	0.047	-2.003‡	0.018	-0.226	0.059	-0.387‡	0.020
Asian other	-0.027	0.133	0.010	0.020	-0.053	0.130	-0.018	0.020
Black Caribbean	-0.137‡	0.048	0.139‡	0.015	-0.093	0.050	0.166‡	0.015
Black African	-0.198‡	0.051	0.063‡	0.015	-0.177‡	0.052	0.100‡	0.015
Black other	-0.187	0.147	0.108‡	0.026	-0.132	0.144	0.149‡	0.028
Chinese	-0.784‡	0.196	0.337‡	0.033	-0.823‡	0.195	0.301‡	0.031
Any other	-0.176	0.128	-0.113‡	0.020	-0.201	0.126	-0.106‡	0.021
Refused	-0.116	0.103	-0.028	0.020	-0.116	0.102	-0.030	0.020
No data	0.017	0.089	-0.241‡	0.019	0.027	0.088	-0.236‡	0.019
Other controls								
Free School Meals	-0.058†	0.027	-0.561‡	0.006	-0.037	0.028	-0.409‡	0.006
Geographic Region	Yes		Yes		Yes		Yes	
Gender × School type	Yes		Yes		Yes		Yes	
Deprived School	Yes		Yes		Yes		Yes	
2001 Census variables	No		No		No		Yes	
Auxiliary parameters								
cut1			-1.441‡	0.012			-0.757‡	0.044
cut2			-1.115‡	0.014			-0.418‡	0.044
cut3			-0.177‡	0.015			0.525‡	0.045
cut4			0.233‡	0.015			0.931‡	0.045
cut5			0.781‡	0.017			1.473‡	0.045

Note: ‡(†) Significant at 1% (5%). OPG standard errors reported. Estimation was performed by Maximum Simulated Likelihood with 800 Halton draws. Adding 100 more draws did not cause important changes on coefficients and/or standard errors.

Table 10. Variables in all equations

Variable	Description	Reason
Benchmark specification		
FSM dummy	Eligible to Free school meal (No)	SES proxy from NPD
Deprived school dummy	Top quintile of %FSM (No)	Design variable
Ethnicity dummies	14 ethnicities (White British)	Variable of main interest; design variable
School-type by gender dummies	4 groups: mixed/boys, mixed/girl, boys/boy, (girls/girl)	Predictor of selection, achievement, mother's education
Geographic region dummies	9 regions (East Midlands)	Predictor of selection, achievement, mother's education
Extra controls from 2001 UK Census at lower layer super output area level		
Population density	persons per hectare	predictor of selection, achievement, mother's education
Index of multiple deprivation	See note	predictor of selection, achievement, mother's education
% population with qualification	6 groups: (no qualifications), other qualifications, GCSE grades A-C, GCE A levels or equivalent, Higher education no degree, Degree or equivalent	predictor of selection, achievement, mother's education
Country of birth	7 groups: (Engalnd), Scotland, Wales, Northern Ireland, Ireland, EU, Any other	predictor of selection, achievement, mother's education

Note. Category in brackets is the reference group. The *index of multiple deprivation* is an indicator that is calculated by the Office of National Statistics and measures deprivation across key themes including income, employment, education and health.More information on the index of multiple deprivation can be found at http://www.neighbourhood.statistics.gov.uk/



(a) Missing covariate (b) Complete data

Circles represent unobserved, or latent, variables whereas rectangles represent observed variables. Arrows connecting latent and/or observed variables represent linear and non-linear relationships. Coefficients are written alongside the relevant arrow. Subfigure (a) depicts the model when the ordinal covariate x is missing, while subfigure (b) depicts the model when the ordinal covariate x is observed.

Figure 1 Path diagram — Missing ordinal covariate with informative selection