Department of Quantitative Social Science

Measuring school value added with administrative data: the problem of missing variables

Lorraine Dearden
Alfonso Miranda
Sophia Rabe-Hesketh

DoQSS Working Paper No. 11-05
June 2011

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.

Measuring School Value Added with Administrative Data: the problem of missing variables

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

Abstract. The UK Department for Education (DfE) calculates contextualised value added (CVA) measures of school performance using administrative data that contain only a limited set of explanatory variables. Differences on schools' intake regarding characteristics such as mother's education are not accounted for due to the lack of background information in the data. In this paper we use linked survey and administrative data to assess the potential biases that missing control variables cause in the calculation of CVA measures of school performance. We find that ignoring the effect of mother's education leads DfE to erroneously over-penalise low achieving schools that have a greater proportion of mothers with low qualifications and to over-reward high achieving schools that have a greater proportion of mothers with higher qualifications. This suggests that collecting a rich set of controls in administrative records is necessary for producing reliable CVA measures of school performance.

JEL classification: I21, C18.

Keywords: contextualised value added, missing data, informative sample selection, administrative data, UK.

[*]Institute of Education, University f London and Institute for Fiscal Studies. E-mail: (l.dearden@ifs.org.uk)

[†]Institute of Education, University of London. E-mail: (A.Miranda@ioe.ac.uk)

[‡]Institute of Education , University of London and University of California, Berkeley. E-mail: (sophiarh@berkeley.edu)

[§]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

As we have already seen in this volume, measuring school effectiveness in a reliable and informative way is very difficult. The previous government's favoured measure, contextualised value added (CVA), looks at the difference between actual outcomes and 'expected' outcomes at various key stages and uses the school average of these differences as a school CVA measure.

'Expected' outcomes are estimated from a model which takes into account prior academic attainment *as well as* other characteristics of the individual (hence it is contextualised) such as age, ethnicity, special educational needs, free school meal status, neighbourhood deprivation, whether the child has English as an additional language and other factors such as school moves. The idea is to measure what is being added *by the school* rather than the impact of factors that happen outside the school (see Raudenbush and Willms, 1995).

But do the contextual variables in the CVA model actually capture everything happening outside the school? In particular does the lack of important family background information in the administrative data lead to biased CVA estimates? This is the subject of this paper. One crucial piece of information that is missing in the administrative data is parental education. Parental education is well-known to affect educational outcomes and is very strongly related to the child's home learning environment (see Blundell et. al. (2006) and Dearden et. al. (2011)). In this paper we use the approach of Miranda and Rabe-Hesketh (2010) to estimate the extent of biases in school CVA measures that are induced by not controlling for mother's education. We find that mother's education varies considerably between schools and that this leads to upward biased estimates of CVA for schools with large proportions of educated mothers and downward biased CVA scores for schools with small proportions of highly educated mothers. It is highly likely that this will also be true for other family background characteristics that proxy what happens in the

home and that systematically differ by school, such as income and family composition[1]. We conclude that CVA measures could be significantly improved if the administrative data included key variables which are likely to proxy what happens in the home, in particular parental education and family composition.

The questions we ask in this paper are as follows. To what extent are current measures of CVA actually reflecting differences in school performance and to what extent are they measuring other unobservable factors not taking place in the school which vary systematically between schools? Are there systematic biases in CVA measures and in what direction? How can these be overcome and what is the appropriate policy response?

2. What is the potential problem with missing data and measuring CVA?

Contextualised value added measures the value added by the school from a baseline test score after all the impacts of factors outside the school are accounted for. In this paper we are going to concentrate on CVA between the ages of 11 (the last year of primary school) and age 16 (the last year of compulsory schooling) which is a key measure for state secondary schools in England. A child's expected outcome at 16 is based on a conditional prediction from a multi-level model with random intercepts for secondary schools which has as explanators/predictors observable characteristics of the child as well as their nationally assessed Key Stage 2 results taken in their last year of primary school. The individual characteristics that are taken into account include things like age within year, gender, ethnicity, English as an additional language, special educational needs, whether child is on free school meals, a measure of neighbourhood deprivation (based on where the child lives) and whether the child joined the school late. Crucially, in the administrative data, there are no measures of family composition, income or parental education which

[1] By family composition we mean things like number of siblings, birth order, gender of sib-ship and whether child lives with both, one or no parents.

vary systematically between schools and are also known to affect educational outcomes.

If an individual has a positive measure of CVA then they are performing better than expected (the school is assessed to be adding value for this individual) and if somebody has a negative CVA then they are performing worse than expected (and the school is assessed as having added negative value for that child). These individual measures are averaged over pupils in the school to get an overall measure of school valued added[2].

CVA is interpreted as the value each school 'adds' over and above what pupils are expected to achieve. It is clearly a *relative* measure and by construction around half of schools will be judged to add value and half to not add value – even if all schools improve markedly from one year to the next. But it is widely used by policy makers, school inspectors and parents to assess the effectiveness of the school. It is also used by teachers and schools to assess progress of individuals as well as for target setting.

To illustrate the potential problem of missing variables in administrative data, we focus on one characteristic which is not measured in English administrative data and that is a well known determinant of academic outcomes – mother's education. From the existing literature we know that parents' education, in particular mother's education, is a significant predictor of children's educational attainment, even after controlling for factors such as income and family size (see, for instance, Crawford et. al., 2010; Haveman and Wolfe, 1995; Behrman, Rosenzweig, and Taubman, 1994). Children of highly educated parents are likely to have access to better learning support at home, and tend to come from families with higher income. As a consequence, they are also more likely to receive private tuition, have parents with higher educational aspirations for their children and the children themselves are

[2] The CVA measure is normalised to have a mean of 1000 rather than zero and there are some adjustments for small schools but the description above is largely accurate.

more likely to have higher educational aspirations. The important aspect of this for measuring CVA is that all this extra support at home is a child and/or family specific characteristic that is outside the control of the school and should not be included in a school's CVA measure. But because parental education, or in our example, mother's education, is not collected in the administrative data, there is a serious risk that its omission from the CVA model will lead to bias in the school CVA measures if mother's education varies considerably between schools. If this is the case, if we do not control for mother's education in the CVA calculations, schools with a large intake of pupils from families with highly (less) educated mothers are likely to be wrongly granted a higher (lower) CVA score than they should.

3. Data and Methodological Approach

In order to explore this problem we exploit the fact that for one cohort of individuals, those born between September 1989 and August 1990, we have a unique data linkage which allows us to assess the implications of missing covariates in the school's administrative data used to calculate school CVA measures.

CVA models are estimated using data from the National Pupil Database (NPD), which has annual basic *individual level* student background information on every child in the state schooling system from January 2002 onwards (Pupil Level Annual Schools Census (PLASC) data). PLASC data can be linked to individual attainment data including outcomes at age 10/11 (Key Stage 2 (KS2) - which has been collected at the individual level since 1996 including some private schools) and 15/16 (Key Stage 4 (KS4) - which has been collected at the individual level since 2002 and includes the private sector).

CVA models are only estimated for the state sector because PLASC data is not available for private schools. The estimates of CVA between KS2 and KS4 are obtained by modelling individuals' KS4 outcome, also known as KS4 Capped Point

Score (CPS), which is a test score that summarizes the results that students obtain in their best 8 subjects. CPS is modelled as a function of:

- Achievement at age 11 (quadratic function of KS2 average point score (APS) in English, Maths and Science and difference between individual English and Maths score and KS2 APS for whole cohort)
- Measures of mobility (pupil joined after September of Year 10 or unusual entry data)
- Other background characteristics (gender, age within year, eligibility for Free School Meals (FSM), Ethnic group, FSM status interacted with Ethnic group, English as an additional language(EAL), EAL interacted with KS2 APS and APS squared, Special Educational Needs (SEN) status)
- Neighbourhood deprivation based on area where child lives (IDACI which measures the proportion of children living in poor households in child's neighbourhood)
- Secondary School intake measures - Average KS2 APS for school and average standard deviation of APS for school

The CVA for each individual is the difference between their actual score and their predicted score from this model. However, for the cohort of individuals born between September 1989 and August 1990 we also have a survey which was drawn from the NPD sample and can be linked to the NPD. This survey, the Longitudinal Survey of Young People in England (LSYPE) or 'Next Steps', first interviewed a sample of English Year 9 pupils in 2004 as well as their parents. Interviews have been conducted annually for a further 6 years. Pupils were sampled using a two stage design. Schools were sampled at the first stage, with over-sampling of schools with high levels of deprivation (measured by FSM). Pupils were sampled from each of the selected schools in the second stage with pupils from some ethnic minority groups being over-sampled to ensure sufficient samples sizes for each ethnicity. Importantly for this paper, the LSYPE measures mother's education (as well as other family background variables like family composition, social class and income). From a total population of 3,117 schools with an average size of 177 pupils, the LSYPE sampled 989 schools with an average sample size of 19 students per school. Mother's education is missing for 4,915 (27%) of these pupils due to unit dropout or item non-response,

but that is explicitly taken into account in our modelling strategy[3]. The sample sizes of the linked data sets are shown in Table 1 below, separating out the LSYPE sample by whether mother's education is recorded or not.

Table 1: Sample sizes in NPD and LSYPE linkage

	NPD	LSYPE		
		Total	Mother's Education Recorded	Mother's Education Missing
Schools	3,117	964	818	774
Pupils	554,320	18,368†	12,978	4,915
Pupils per school	177	19.1	15.9	6.4

† For 475 of these cases, mother's education is missing because the mother was reported to be "not a member of the household" but survey was otherwise completed.

Further details of mother's education for whom we have information is given in Table 2. In our modelling, we consider 4 levels of mother's education: no qualifications, GCSE qualifications or below (Level 1/Level 2 qualifications); A level qualifications or equivalent (Level 3 qualification); Degree or higher qualifications (Level 4/Level 5 qualifications).

Table 2: Mother's Highest Qualification, mean KS4 Capped Point Score (CPS) and mean KS4 CVA score in standard deviation units[1]

		Unweighted			Weighted[2]		
Mothers' Highest Qualification Level	Obs.	%	KS4 CPS	KS4 CVA	%	KS4 CPS	KS4 CVA
1. No qualifications	3,397	26.2	-0.252	-0.054	19.6	-0.447	-0.176
2. Level 1/Level 2	5,178	39.9	-0.002	0.002	44.3	-0.047	-0.039
3. Level 3	3,181	24.5	0.296	0.096	26.2	0.289	0.079
4. Level 4 /Level 5	1,222	9.4	0.674	0.256	9.9	0.685	0.252
Total	12,978						

[1] The standard deviation of the KS4 CPS is 67.3.
[2]Weighted using inverse probability of selection weights, adjusted for nonresponse.

The table also reports mean KS4 capped point scores (CPS) and CVA scores that have been standardized by dividing by the standard deviation of the KS4 CPS, so that they can be interpreted in standard deviation units. From this summary table we see that, as one

[3]Mother's education may be missing in the LSYPE for various reasons, including survey unit and item non response. For 475 of these cases, mother's education is missing because the mother was reported to be "not a member of the household" but survey was otherwise completed.

would expect, the standardised KS4 capped point score (CPS) increases with mother's education. However, the story does not end there. Table 2 also shows that the mean standardised CVA score also increases with mother's education. In fact, pupils with mothers in the top qualification category (level 4 / level 5) score on average 0.3 standard deviations higher than pupils with mothers in the bottom category, corresponding to a difference of about 20 CVA points. These differences are highly significant (at the 1% level). This clearly shows that the control variables that enter the official CVA calculations are not sufficient to proxy mother's education. In fact, only 26% of the variance in mother's education is explained by the covariates in the CVA model.[4] As a consequence, CVA scores that do not take into account mother's education are likely to overstate the real value added by schools with high proportions of highly educated mothers and underestimate the real value added by schools with low proportions of highly educated mothers.

In order to assess how much omission of mother's education from the CVA model biases CVA scores, we will construct CVA scores that control for mother's education and compare them with the CVA scores from the DfE. Since mother's education is missing for those not sampled into the LSYPE and for about 30% of the LSYPE sample who did not participate in the survey or did not provide information on mother's education, we will use the methods for handling missing ordinal covariates introduced by Miranda and Rabe-Hesketh (2010).

We start with a brief nontechnical account of our approach and then provide the technical details. Although our approach takes into account that individuals who provide information on mother's education may not be representative of the entire LSYPE sample, we initially ignore this aspect of the model for simplicity. We treat both mother's education and KS4 CPS as dependent variables and model them simultaneously using the merged data. Mother's education is regressed on

[4] Using an ordinal probit regression, 26% of the variance of the latent responses is explained by the covariates.

covariates from the NPD (that are available for everyone), and KS4 CPS is regressed on mother's education (sometimes unobserved) and covariates from the NPD. Information on the relationship between mother's education and variables from the NPD comes from the subsample of pupils for whom mother's education is recorded, but this relationship is assumed to be the same for all pupils. The entire population then provides information on the relationship between KS4 CPS and mother's education, controlling for the other variables. Those for whom mother's education is missing provide this information only indirectly, through the relationship between mother's education and the NPD covariates, taking into account that mother's education is not perfectly predicted by the NPD covariates. To correct for sample selection bias, our model treats selection as a third dependent variable that is allowed to be correlated with mother's education and KS4 CPS, similar to a Heckman selection model.

It is straightforward to obtain the 'expected' outcomes needed to calculate CVA scores when mother's education is recorded, When mother's education is missing, we use the pupil-specific 'posterior' probabilities of the different levels of mother's education (posterior meaning that we take into account all available data) to obtain the 'posterior mean' expected outcomes. We also obtain standard errors that take into account the uncertainty regarding the true level of mother's education and are therefore larger than the standard errors for pupils for whom mother's education is known.

We now give some technical details. As we just said, the objective is to control for mother's education (x_i) in the model for the KS4 capped score (y_i), and handle missing values of mother's education through joint modelling of y_i and x_i (see also Little and Schluchter, 1985). We know that x_i is missing at random for pupils who were not sampled into the survey, but we correct for selection bias due to survey/item non-response among sampled pupils ($S_i = 1$ if mother's education is recorded, $S_i = 0$ if mother's education not recorded). This is accomplished by using a simultaneous model for y_i, x_i, and S_i that allows selection S_i to be associated with

mother's education x_i and pupil achievement y_i, after controlling for the other explanatory variables. (For related approaches, see Heckman, 1979; Lipsitz et al., 1999; Wu and Carroll, 1988.).The methodological approach thus explicitly takes into account that survey/item response in the LSYPE is likely to be endogenous or non-ignorable (see also Heckman, 1979). Miranda and Rabe-Hesketh (2010) show that mothers of high achieving children are more likely to respond, and highly educated mothers are less likely to respond. Of course, mother's education is missing if the individual is not in the LSYPE ($S_i = .$) but we assume that if we control for LSYPE sample design variables (that determine the sampling probability namely region, FSM status and ethnicity), this missingness will be ignorable (Rubin, 1987).

In order to identify the model, we need credible exclusion restrictions and we need to predict mother's education based on covariates observed in the NPD sample. In our model we regress KS4 capped score on the traditional covariates (w_i) entered in the DfE CVA model described earlier as well as mother's education. To identify the model we assume that one of the CVA covariates in w_i, the student's relative age within year, determines KS4 CPS (y_i) but not mother's education (x_i) or selection (S_i). Relative age is well-known to affect children's academic outcomes in England (see Crawford et. al. 2010), but should have no impact on mother's education or item and survey non-response. We also assume that some of the covariates (r_i) in the selection equation, in particular dummy variables identifying the different LSYPE interviewing contractors (4 different contractors), affect selection (S_i) but not mother's education (x_i) or KS4 CPS (y_i). In the LSYPE survey there are marked and significant differences in whether mother's education is recorded depending on the contractor who is undertaking the survey. The covariates in the mother's education equation (z_i) are the same as for the selection equation (gender, ethnicity, English as an additional language, special educational needs, whether child is on free school meals, a measure of neighbourhood deprivation, regional dummies) but exclude the contractor dummies. We allow for correlation between the errors of S_i and y_i and S_i and x_i due to unobserved characteristics.

More formally, the variable of interest, KS4 CPS (y_i), is modelled as a linear regression as shown in equation (1):

$$y_i = \underbrace{\sum_{g=1}^{4} 1(x_i = g)\beta_g}_{\text{latent when } x_i \text{ missing}} + w_i'\theta + \varepsilon_{yi} \quad (1)$$

where w_i are the covariates contained in the DfE CVA equation, $x_i=1$ if the mother has no qualifications, $x_i=2$ if the mother has level1/level2 qualifications, $x_i=3$ if the mother has level 3 qualifications and $x_i=4$ if the mother has level4/level5 qualifications and ε_{yi} is the error term. Mother's education is unknown, or latent, for the majority of pupils. Hence the main methodological challenge is to use a discrete latent variable η_i in place of the first term in (1) when S_i is missing or $S_i=0$. Such a latent variable should comply with the property that, given observable characteristics, the probability that $\eta_i=\beta_g$ in the sample with missing x_i should be equal to the probability that $x_i=g$ in the sample where x_i is actually observed. We use an ordered probit model for this probability, written in terms of a latent response x_i^* as shown in equation (2):

$$x_i^* = z_i'\gamma + \varepsilon_{xi} \quad (2)$$

Here $x_i=g$ if $\kappa_{g-1} < x_i^* < \kappa_g$, $g=1,2,3,4$, $\kappa_0=-\infty$, $\kappa_4=\infty$ and κ_1, κ_2, κ_3 are threshold parameters. The covariates z_i are as described above and come entirely from the NPD data, γ are regression coefficients, and ε_{xi} is an error term. Finally, selection is modelled as a binary probit with the model for the latent response S_i^* as shown in equation (3)

$$S_i^* = r_i'\alpha + \varepsilon_{si} \quad (3)$$

Here $S_i=1$ if $S_i^*>0$, r_i are the same covariates as in the mother's education equation but also include survey contractor dummies, α are regression coefficients, and ε_{Si} is

an error term. Correlation between the errors in the equations for S_i^* and y_i and S_i^* and x_i^* due to unobserved characteristics are explicitly modelled. If these correlations are different from zero we say that there is 'informative selection.' Informative selection may arise, for instance, because mothers of high performers are more interested on their children education and so more likely to respond to the LSYPE survey (in which case corr(ε_{yi}, ε_{si})>0). Or because highly educated mothers are more difficult to interview because they have busy professional jobs and, therefore, less available for contact (in which case corr(ε_{xi}, ε_{si})<0).

The model is estimated using maximum simulated likelihood methods. The estimates for the equation for KS4 CPS are shown in Table 3 below. For interpretability, the KS4 CPS was rescaled to have a standard deviation of 1 but for the purpose of constructing CVA scores, regression coefficients were transformed to the original scale. For comparison, estimates from a simple random effects linear model (without mother's education) are reported along those obtained from the missing covariate model. From the bottom of the table, we can see that mother's education plays an important role in explaining KS4 CPS. (There is no reference category here so differences in mean standardized KS4 CPS between levels of education, controlling for the other variables, are differences between the corresponding coefficients.)

Table 3: Standardised capped new style GCSE score equation. ‡ (†) Significant at 1% (5%). (a) Model does not contain intercept; Outer Product Gradient (OPG) standard errors reported (see Berndt et. al. 1974),

Variable	Linear model		Missing Covariate Model[(a)]	
	Coefficient	SE	Coefficient	SE
Pupils characteristics				
Standarised KS2 average point score (sd_KS2)	0.586‡	0.001	0.542‡	0.001
Sd_KS2 squared	0.053‡	0.001	0.042‡	0.001
Difference between sd. English score and Sd_KS2	0.097‡	0.003	0.098‡	0.002
Difference between sd. Mathematics score and Sd_KS2	0.018‡	0.003	0.032‡	0.002
IDACI	-0.614‡	0.007	-0.403‡	0.004
Special education needs-school action	-0.348‡	0.003	-0.225‡	0.002
Special education needs statement plus school action	-0.599‡	0.004	-0.371‡	0.002
Mover year 10	-0.732‡	0.007	-0.492‡	0.004
Mover within year	-0.229‡	0.004	-0.154‡	0.003
Female	0.136‡	0.002	0.107‡	0.001
Age within year	-0.133‡	0.003	-0.103‡	0.002

Table 3 (continued)

Variable	Linear model		Missing Covariate Model[(a)]	
	Coefficient	SE	Coefficient	SE
Pupils characteristics				
English as additional language (EAL)	0.183‡	0.006	0.169‡	0.004
EAL interacted with sd_KS2	-0.107‡	0.004	-0.101‡	0.002
EAL interacted with sd_KS2 squared	-0.023‡	0.002	-0.004‡	0.001
Free school meals (FSM)	-0.236‡	0.003	-0.082‡	0.002
White other	0.077‡	0.007	0.110‡	0.005
Mixed	0.031‡	0.007	0.067‡	0.004
Indian	0.229‡	0.008	0.252‡	0.005
Pakistani	0.172‡	0.009	0.280‡	0.005
Bangladeshi	0.235‡	0.015	0.754‡	0.01
Asian other	0.252‡	0.014	0.243‡	0.01
Black Caribbean	0.106‡	0.009	0.103‡	0.005
Black African	0.264‡	0.011	0.256‡	0.006
Black other	0.062‡	0.016	0.053‡	0.012
Chinese	0.306‡	0.017	0.274‡	0.012
Other ethinic group	0.149‡	0.013	0.201‡	0.009
Refused ethinicity question	-0.057‡	0.009	-0.021‡	0.007
No data in ethnicity field	-0.127‡	0.009	-0.039‡	0.006
FSM interacted with White other	0.209‡	0.017	0.199‡	0.011
FSM interacted with Mixed	0.071‡	0.014	0.097‡	0.008
FSM interacted with Indian	0.177‡	0.018	0.129‡	0.01
FSM interacted with Pakistani	0.203‡	0.012	0.244‡	0.008
FSM interacted with Bangladeshi	0.21‡	0.018	0.751‡	0.011
FSM interacted with Asian other	0.243‡	0.03	0.203‡	0.019
FSM interacted with Black Caribbean	0.194‡	0.017	0.140‡	0.01
FSM interacted with Black African	0.198‡	0.017	0.192‡	0.01
FSM interacted with Black other	0.220‡	0.03	0.198‡	0.021
FSM interacted with Chinese	0.275‡	0.045	0.148‡	0.032
FSM interacted with other ethnic group	0.261‡	0.022	0.194‡	0.015
FSM interacted with refused ethnicity question	0.047‡	0.024	0.043‡	0.017
School characteristics				
FSM interacted with no data in ethnicity field	0.039‡	0.024	0.064‡	0.016
Average sd_KS2 of intake for school	0.099‡	0.011	0.072‡	0.003
Average standard deviation of sd_KS2 of intake for school	-0.183‡	0.029	-0.158‡	0.007
Mother's education				
1. No qualifications			-1.175‡	0.007
2. Level 1/Level 2			0.355‡	0.007
3. Level 3			0.384‡	0.008
4. Level 4 /Level 5			0.438‡	0.009
Corr(S,y)			0.147‡	0.010
Corr(S,x)			-0.294‡	0.015
Var(ε_y)			0.232‡	0.000
N. observations	544,175		544,175	

To obtain revised CVA measures, we need to predict the capped GCSE score for everybody in the NPD sample. If the individual is in the LSYPE and mother's education is measured then our predicted score is simply given by equation (4):

$$\hat{y}_i = \sum_{g=1}^{G} 1(x_i = g)\hat{\beta}_g + w_i'\hat{\theta} \qquad (4)$$

For those whose mother's education is missing, our predicted score is given by equation (5)

$$\hat{y}_i = \sum_{g=1}^{G} P_{gi}(\hat{\beta}_g + w_i'\hat{\theta}) = \sum_{g=1}^{G} P_{gi}\hat{\beta}_g + w_i'\hat{\theta} \qquad (5)$$

For non-responders in the LSYPE sample, the required posterior probability P_{gi} is given by

$$P_{gi} \equiv P(g \mid y_i, S_i = 0) = P(x_i = g, y_i, S_i = 0) / \sum_{c=1}^{G} P(x_i = c, y_i, S_i = 0).$$

For people who were not sampled into LSYPE (i.e. S_i is missing), the posterior probability is given by

$$P_{gi} \equiv P(g \mid y_i) = P(x_i = g, y_i) / \sum_{c=1}^{G} P(x_i = c, y_i).$$

Once we have these predicted scores for our whole sample we can calculate our corrected CVA measures and standard errors for these measures. Our CVA measure for school *j* with n_j pupils is simply given by:

$$CVA_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij}) \qquad (7)$$

4. Results

In this section we compare our missing covariate adjusted (MCA) measure of CVA with the DfE measure. In Figure 1, we show a scatter plot with percentiles based on our MCA-adjusted measure on the vertical axis and percentiles based on the DfE measure on the horizontal axis. If both measures were broadly the same, all points on the scatter plot would lie around the 45-degree line. However, we see many points in the top-left where ignoring the effect of mother's education on school achievement leads DfE to erroneously over-penalize low achieving schools, and on

the bottom-right where high-achieving schools (who have many more mothers with higher qualifications) are over-rewarded.

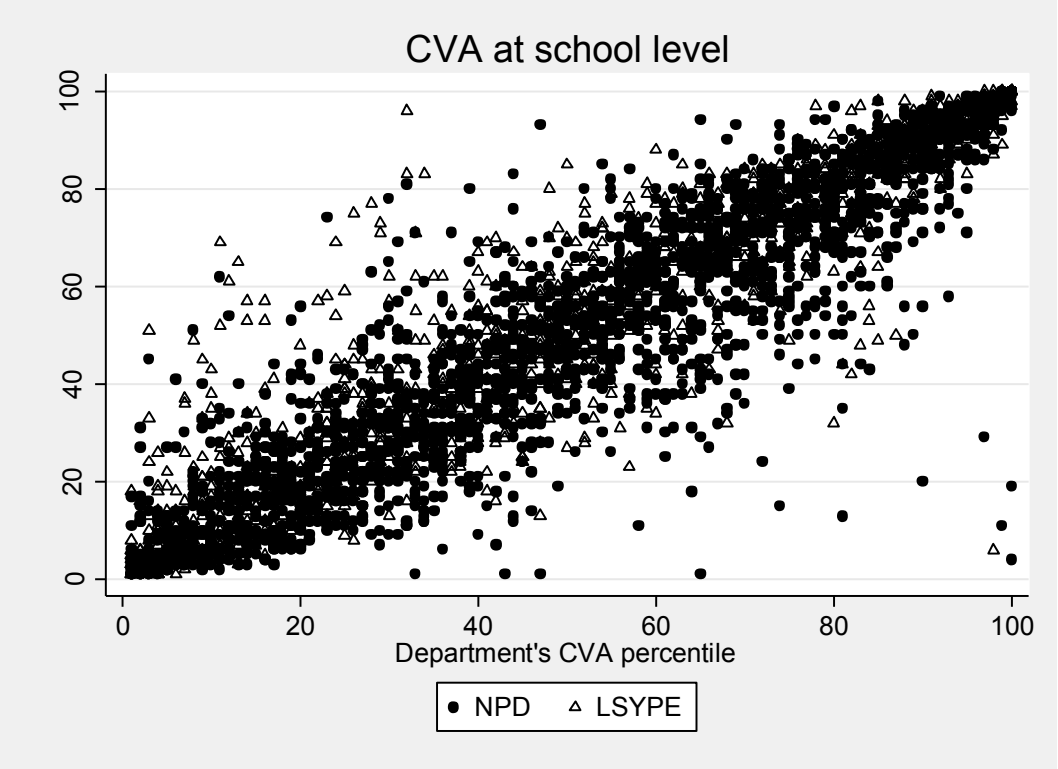

Figure 1. Scatter Plot of percentiles based on school level MCA CVA and DfE CVA measures

Standard errors for pupil-level MCA CVAs were estimated to reflect the imprecision of the estimated regression coefficients and, if not observed, the uncertainty regarding mother's education[5]. Similarly, standard errors for pupil-level DfE CVAs were estimated to reflect the imprecision of the estimated regression coefficients. Under the conservative assumption that the errors in the two measures of CVA are independent of each other, we tested the difference in pupil-level CVA measures and found significant differences, at the 5% level, for 46% of the pupils for whom mother's education was unknown and for 95% of pupils for whom mother's

[5]For details on how to calculate standard errors for the predicted y_i see Miranda and Rabe-Hesketh (2010)

education was known. Aggregating pupil-level CVA measures to the school level (under the conservative assumption that the errors in the measures are independent across pupils), gave very few significant differences because standard errors were large due to the many students per school whose mothers' education was unknown. To better reflect the hypothetical situation of collecting mother's education for all pupils, we constructed both the DfE CVA and the adjusted CVA measures at the school-level by averaging the results of only those pupils whose mother's education was known. We restricted our comparison to the 614 schools for which mother's education was known for at least 10 pupils. On average, 22 pupils per school contributed to the CVA scores. We found significant differences between the CVA measures for 94% of these schools (at the 5% level).

Conclusions

In this paper, we use unique linked survey and administrative data to assess the possible biases in the currently used DfE CVA measure caused by not controlling for mother's education. We use this as an illustration of the potential problems caused by not having rich controls in administrative data used to estimate these models. At present in the UK data we have good measures of important factors like ethnicity and poverty (measured by eligibility for free school meals) but we don't observe factors such as parental education and family size which are well known determinants of educational outcomes. If these unobserved variables are proxies for what is happening in the home rather than the school, and if they differ systematically by school, then the potential for large biases in CVA measures for schools with disproportionately high or low levels of these missing variables will be large.

We have not addressed other methodological concerns regarding the DfE CVA measure. In particular, the regression coefficients are likely to be biased because pupils' family background is likely to be correlated with the school value-added (and other school-level variables), but the DfE approach, as well as our approach, assume

that they are not. Raudenbush and Willms (1995) address this problem by using the consistent fixed-effects estimator of the regression coefficients. We did not adopt this approach here because we wanted to focus on the problem of omitting important backround information from the model, keeping other aspects of the analysis as similar as possible to the DfE method.

The policy response to the problem identified in this paper is reasonably simple: to collect better background information in the PLASC data. There is a large literature on the factors that impact on educational outcomes. Some of this would be impossible to collect in administrative data (such as family income). But other important determinants, such as parental education and family size (e.g. how many older and younger siblings each child has), could also be collected as part of the PLASC return. This seems feasible given that parents already provide some information about their children to the schools, such as ethnicity. Rectifying this omission would, without doubt, lead to fairer and more meaningful measures of school value added as well as increasing the value and use of English schools administrative data more generally.

5. References

Behrman, J., Rosenzweig, M. and Taubman, P. (1994), 'Endowments and the Allocation of Schooling in the Family and in the Marriage Market: The Twins Experiment', Journal of Political Economy, vol. 102, pp. 1131-74.

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

Blundell, R., Dearden, L. and Sianesi, B. (2005), 'Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey', *Journal Of The Royal Statistical Society Series A*, vol. 168, pp. 473-512.

Crawford, C., Dearden, L. and Meghir, C. (2010), 'When you are born matters', DoQSS Working Paper No. 10-09.

Dearden, L., Sibieta, L. and Sylva, K. (2011), 'The socio-economic gradient in early child outcomes: evidence from the Millennium Cohort Study', *Longitudinal and Life Course Studies*, vol. 2, pp. 19-40.

Haveman, R. and Wolfe, B. (1995), 'The Determinants of Children's Attainments: A Review of Methods and Findings', *Journal of Economic Literature,* vol. 33, pp. 1829-1878.

Heckman, J.J. (1979), 'Sample selection bias as a specification error', *Econometrica,* vol. 47, pp. 153-161.

Leckie, GB. and Goldstein, H. (2009), 'The limitations of using school league tables to inform school choice', *Journal of the Royal Statistical Society (Series A),* vol. 172, pp. 835-851.

Lipsitz, S.R., Ibrahim, J.G., Chen, M.H. and Peterson, H. (1999), 'Non-ignorable missing covariates in generalized linear models', *Statistics in Medicine,* vol. 18, pp. 2435-2448.

Little, R. J.A. and Schluchter, M. (1985), 'Maximum likelihood estimation for mixed continuous and categorical data with missing values', *Biometrika,* vol. 72, 497-512.

Raudenbush. S. W. And Willms, J. D. (1995), 'The Estimation of School Effects'. *Journal of Educational and Behavioral Statistics,* vol. 20, pp. 307-335.

Ray, A. (2006), 'School value added measures in England, Tech. rep., Department for Education and Skills. Document available at: http://www.dcsf.gov.uk/research/data/uploadfiles/RW85.pdf

Rubin, D.B. (1976), 'Inference and missing data', *Biometrika*, vol. 63, pp. 581-592.

Miranda, A. And Rabe-Hesketh, S. (2010), 'Missing ordinal covariates with informative selection', DoQSS Working Papers No. 10-16.

Wu, M.C. and Carroll, R.J. (1988), 'Estimation and comparison of change in the presence of informative right censoring by modeling the censoring process', *Biometrics,* vol. 44, pp. 175--188.