Methodological Advice on Family Occupation and Education Index

Dr Carole Birrell and Professor David Steel

Centre for Sample Survey Methodology

National Institute for Applied Statistics Research

University of Wollongong

Summary

The Family Occupation and Education Index (FOEI) is a school-level socio-economic index developed by the NSW Department of Education and Communities (DEC). The FOEI is based on school-level regression analysis of the relationship between the average of standardised students' achievement scores obtained from NAPLAN results and parental background variables, which are the level of highest school education, highest non-school education and occupation. In developing the 2013 FOEI, the National Institute for Applied Statistics Research Australia at the University of Wollongong has been contracted to review and provide external validation of the FOEI methodology due to the proposed use of the FOEI data in resource allocation.

For the purpose of this review, a sample of 2012 data was provided by NSW DEC to NIASRA. The sample data included approximately 50% of schools with all students at the selected schools. Data included, for each de-identified student record, student year of schooling, gender, Aboriginal status, reported and standardised NAPLAN reading and numeracy results, parent education and occupation variables, and a set of school and community variables derived from the 2011 ABS census.

In developing the 2013 FOEI several technical issues associated with missing data and regression analysis have been considered and the following approaches are recommended to cope with them.

Imputation

A. Use a model-based multiple imputation approach to reduce bias in FOEI arising from missing parental background data

Data on parents can be missing. This missing data affects the estimation of the regression function and the calculation of FOEI scores for individual schools. A model-based multiple imputation approach – multiple imputations by chained equations (MICE) – is recommended to deal with missing data. This approach uses the relationships between the variables in the observed data to impute plausible values for the missing data. This is done multiple times (M=10) to enable valid estimates of uncertainty accounting for both the regression model estimation and the imputation itself. This is a widely adopted and flexible approach that allows the full use of the observed data for many variables and can be implemented using readily available statistical software (White et al, 2011). As with any imputation approach the method is based on some assumptions, including the assumption that conditional on the observed data the unobserved data are missing at random. The missing at random (MAR) assumption is less restrictive than the assumption of Missing Completely at Random (MCAR) which is assumed if complete cases, that is students for whom all parental data are available, only are used in the analysis. It is possible that the mechanism of missingness is Missing Not at Random (MNAR), however, the use of a large number of explanatory variables in the imputation model should assist in moving closer to the MAR assumption and therefore reduce any possible bias (White et al, 2011).

B. *Use as many relevant variables as possible to help the imputation process*

It is recommended that the parental variables are used along with the standardised student achievement scores (when available), ATSI status, school remoteness and a set of community variables derived from the 2011 ABS census including levels of education and occupation of persons in the same statistical area (SA1) as the student's address. These community census variables consist of percentages of people/families and were calculated for all people or all families in the same area as the student's address.

The imputation does not explicitly use the school indicator variable, although the imputation uses the community variables and therefore reflects some characteristics of the local community. Imputation taking into account the nesting structure of the data (i.e., students are clustered in schools) was not adopted because of concerns about the stability of relationships based on small numbers of responding cases in many schools. Some of the larger schools would have had sufficient cases to consider imputation within the school, but then the models of imputation would have differed across schools, which was considered unreliable.

C. *Use different imputation models for one parent and two parent students*

Some students have one parent and some have two parents in the data file. To allow for the use of information on the second parent, the imputation is conducted separately for cases where there are data for two parents and cases with data for only one parent. The imputation for one parent and two parent cases was done separately, on the basis of the department's research that shows the relationship between the parent background variables is different for the one parent variables and the equivalent variables in the two-parent data file. There is a sufficient number of students from both one and two parent families to enable the relationships amongst parent variables to be estimated separately for the two groups of students.

D. *Use multiple years of NAPLAN data to maximise the number of students with available standardised achievement data to help the imputation process*

For example, in the sample data provided, standardised NAPLAN reading and numeracy results for students in Years 3, 5, 7 and 9 in 2012 were used in addition to matched 2011 results for students in Years 4, 6, 8 and 10 in 2012. This allowed standardised achievement data to be used, in the imputation process for the review, for Years 3 - 10 for students in 2012.

E. Use a different imputation model for students without NAPLAN achievement scores

Students who were exempted from NAPLAN or from cohorts for which NAPLAN data does not exist in the review data set (e.g. students in Kindergarten, Year 1, and Year 2) are included in the imputation with other students but student achievement scores are unavailable and therefore not used in the imputation.

Regression Analysis

F. The FOEI regression model proposed by DEC can be implemented using the following approaches:

- The dependent variable in the regression analysis is based on the most recent observed student achievement scores. For the 2013 FOEI, these are the 2012 NAPLAN reading and numeracy scores, standardised and averaged, for students now in Years 4, 6, 8 and 10 in 2013.
- The explanatory variables are the school level parental background variables (i.e. the percentages of parents in each category of school education, non-school education and occupation). These are calculated using parental background data for all students in the school in 2013, including the imputed values for missing data.
- The regression analysis uses a dependent variable based on observed 2012 student achievement data for students in Year 4, 6, 8 and 10 in 2013, while the explanatory variables are based on parental background variables for all students attending the school in 2013. The regression equation is designed to reflect the relationship between achievement in the most recent year for which achievement data are available and the parental background of students. Since calculation of the FOEI score for a school is based on the parental background of all students, estimation of the regression function also uses these as the explanatory variables.
- The review considered the inclusion of ABS community variables in the regression analysis. Previous regression analysis performed by the DEC suggested no appreciable additional predictive power is added if the community level variables are used in the school-level regression analysis. However, community level variables are useful information to help impute for missing parental data.

G. Weighting parental information for students in one parent families

In the regression estimation and in the production of the FOEI score the school-level parental variables effectively average the characteristics of the parents in a two-parent case. In order for each student's family to count equally towards the school FOEI score, the calculation of the school-level variables assigns a weight of 2 to single parents and a weight of 1 to each parent in a two-parent family.

H. Robust regression is the regression technique recommended to construct FOEI scores to deal with outliers

Analysis of residual patterns from ordinary least squares (OLS) and comparisons of OLS, weighted least squares (WLS) and robust regression models show that robust regression is a technique that is effective in reducing the influence of outliers on the regression estimates. The majority of outliers from the OLS analysis are selective schools and small schools. However not all small schools are outliers; in fact the majority of small schools fit the regression model reasonably well. Robust regression reduces the weight for the outliers, i.e., schools that have large residuals. Using robust regression appears to adequately account for both selective schools and the small schools with large residuals, and is recommended.

1. Introduction

The Family Occupation and Education Index (FOEI) is a school-level socio-economic index developed by the NSW Department of Education and Communities (DEC). The FOEI is based on school-level regression analysis of the relationship between the average of standardised students' achievement scores obtained from NAPLAN results and parental background variables, which are the level of highest school education, highest non-school education, and occupation. In developing the 2013 FOEI, the National Institute for Applied Statistics Research Australia (NIASRA) at the University of Wollongong was contracted to review and provide external validation of the FOEI methodology due to the proposed use of the FOEI data in resource allocation.

For the purpose of this review, a sample of 2012 data was provided by NSW DEC to NIASRA. The sample data included approximately 50% of schools with all students at the selected schools. Data included, for each de-identified student record, student year of schooling, gender, Aboriginal status, reported and standardised NAPLAN reading and numeracy results, parent education and occupation variables, and a set of community variables derived from the 2011 census. In developing the 2013 FOEI several statistical issues associated with regression analysis and imputation of missing data have been considered and approaches identified to cope with them.

In section 2 the major methodological issues identified with the regression analysis are considered. In section 3 major methodological issues identified with the imputation for missing data are considered. In section 4 aspects of the interaction between the imputation methodology and the regression analysis are discussed.

The review focussed on the main issues that were considered important for the calculation of the 2013 FOEI scores. Not all issues identified were examined in detail or empirical analysis conducted because of the limited time, data and resources available. However these issues are listed, as they may be examined in more detail in the future.

2. Regression Modelling

In examining the approach to take to the regression analysis the following issues were considered:

- Basics of school-level modelling
- Implications and interpretation of school-level modelling
- Use of robust regression methods
- Use of school achievement data
- Students with one parent and students with two parents
- Use of community variables and other variables in the regression model.

2.1 Basics of School-level Modelling

The basic statistical model underpinning the FOEI is

$$\overline{y}_g = \overline{\mathbf{x}}_g^T \beta + e_g \quad (1)$$

where for school *g*;

- $\overline{y}_g$ is the average achievement measure for all students in the school,
- $\overline{\mathbf{x}}_g$ is the vector of means of the explanatory variables calculated for all students in the school,
- e_g is an error or residual term.

At the school level the explanatory variables are the proportions of parents in each category of highest school education, highest non-school education, and occupation.

The school-level achievement measure is calculated by calculating a standardised score for reading and numeracy for each student and then calculating a simple average of the two standardised scores for each student and then averaging these scores across all the students in the school. For a particular year the scores are only available for students in years 3, 5, 7 and 9. The state government schools mean and standard deviation for the relevant calendar year, grade cohort, and test domain, is used in the standardisation.

If $\hat{\beta}$ is an estimate of the vector of regression parameters then the estimated fitted value for school *g* is

$$\tilde{y}_g = \overline{\mathbf{x}}_g^T \hat{\beta} \quad (2)$$

A simple approach is to estimate β using ordinary least squares (OLS), which will be denoted by $\hat{\beta}_{OLS}$. Provided $E\left[e_g \mid \overline{\mathbf{x}}_g \right] = 0$ then $\hat{\beta}_{OLS}$ is unbiased for β. Also, if the variance of the error term is constant, $V(e_g \mid \overline{\mathbf{x}}_g) = \sigma^2$, then $\hat{\beta}_{OLS}$ is efficient in the sense that it has minimum variance. If the variance is not constant, so $V(e_g) = \sigma_g^2$, then $\hat{\beta}_{OLS}$ is still unbiased but not fully efficient and a weighted least squares estimate, $\hat{\beta}_{WLS}$, which uses weights inversely proportional to σ_g^2 can be considered. In this case the residuals will have different dispersion for different values of the regression term $\overline{\mathbf{x}}_g^T \beta$.

The FOEI for a school is the fitted value, which is the expected school-level achievement based on the parent variables used. The estimated residual term is $\hat{e}_g = \overline{y}_g - \tilde{y}_g$, which reflects how much the school level achievement is above or below the expected value for a school with the values of the parent variables.

2.2 Implications and Interpretation of School-Level Regression

An important feature of model (1) is that it is a school-level model based on data aggregated over all the students in the school. Such models are sometimes called aggregate or ecological models. There are two consequences of the use on an aggregate model that need to be considered: the interpretation of the regression term and the structure of the variance of the error or residual term.

In general the estimated regression coefficient obtained from an aggregate regression model will be different from the estimate that would be obtained from a corresponding analysis of the unit level, in this case student-level, data. This difference is sometimes referred to as the ecological fallacy or ecological bias. The source of this effect can be explained by considering a multilevel regression model for student-level data:

$$y_{ig} = \mathbf{x}_{ig}^T \beta_I + \overline{\mathbf{x}}_g^T \beta_C + u_g + \varepsilon_{ig} \tag{3}$$

where for student *i* in school *g*;

- y_{ig} is the achievement measure
- $\mathbf{x}_{ig}$ is the vector of explanatory variables
- ε_{ig} is an individual level error or residual term
- u_g is a school-level error or residual term

The elements of the vector of explanatory variables take the values one or zero depending on whether the parent belongs to the relevant category or not. The model still includes $\overline{\mathbf{x}}_g$, which can be regarded as contextual effect and is included so that the impact of the proportion of parents in the various categories can be reflected in the regression model.

Model (3) includes the direct effect of a student's parent variables on their achievement measure, with regression coefficient β_I and the indirect effect of the school context as reflected in the school means of the explanatory variables, with regression coefficient β_C. It also includes an error or residual term at the student and school-level. The school level residual reflects the impact of unmeasured factors that affect the achievement measure of all students in a school that produce a degree of similarity in their achievement.

Aggregating model (3) leads to

$$\overline{y}_g = \overline{\mathbf{x}}_g^T \beta_{I+C} + u_g + \overline{\varepsilon}_g \tag{4}$$

where $\beta_{I+C} = \beta_I + \beta_C$ and $\overline{\varepsilon}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \varepsilon_{ig}$. Model (4) is of the same form as model (1) where $\beta = \beta_{I+C}$ and $e_g = u_g + \overline{\varepsilon}_g$.

A standard individual (i.e. student) level regression analysis will effectively assume that $\beta_C = 0$ and $\beta_I = \beta$. The estimates of the regression coefficients from an aggregate (i.e. school level) and unit (i.e. student) level analysis will have the same expectation if $\beta_C = 0$ otherwise they will differ, which is the ecological fallacy. However, this is not a concern for the FOEI since the aggregates analysis will produce unbiased estimate of β_{I+C}, which is appropriate.

The aggregate model (4) that arises from (3) makes it clear that the school-level analysis is reflecting the combined effect of student-level explanatory variables and their contextual effects.

A consequence of model (4) is the variance of the school mean measure achievement based on n_g students is

$$V(\bar{y}_g \mid \bar{\mathbf{x}}_g) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n_g} \tag{5}$$

Essentially, because the school measure is a mean it will have a component of variance that is a function of the inverse of the sample size in the school. Hence, we should expect more variation in the school-level measure for smaller schools.

Estimation of σ_u^2 and σ_ε^2 would usually be done using unit level, that is student level data and using the school indicator. Given the likely existence of contextual effects, even a unit level model should include the contextual effects $\bar{\mathbf{x}}_g^T \beta_C$. A simple diagnostic that can be useful in giving some indication of the relative size of the variance component can be based on analysis of the school-level residuals $\hat{e}_g$, such as plotting $\hat{e}_g^2$ against n_g or n_g^{-1}.

The form of (5) suggests that OLS will be close to efficient if σ_u^2 is much larger than $\frac{\sigma_\varepsilon^2}{n_g}$. Alternatively if σ_u^2 is much smaller than $\frac{\sigma_\varepsilon^2}{n_g}$ then a weighted least squares (WLS) regression estimates, with weight n_g is more appropriate. We can denote this estimate as $\hat{\beta}_{WLS}$. This suggests a useful diagnostic is to run OLS and WLS analyses. A formal multilevel modelling approach will iteratively estimate the regression coefficient and variance component using model (3). It will estimate the regression coefficient efficiently and produce valid estimates of standard errors. To avoid unit level modelling we can note that (5) can be rewritten as

$$V(\bar{y}_g \mid \bar{\mathbf{x}}_g) = \sigma^2 \left(\frac{1-\rho}{n_g} + \rho \right) \tag{6}$$

Here $\rho = \sigma_u^2 \big/ \left(\sigma_u^2 + \sigma_\varepsilon^2 \right)$ is the intra-class correlation, which is the correlation between the achievement measure for students in the same school due to the unmeasured factors reflected in u_g. A WLS school-level analyses can then be

carried out using weights obtained as the inverse of (6) for a reasonable value of ρ, perhaps obtained from analysis of the school-level residuals.

In practice the different weights that can be used may not affect the estimates of the regression coefficients appreciably, although they may affect the estimated standard errors (SEs). However, recognising the variance structure reflected in (5) can be important for identification of outliers and in producing confidence intervals associated with the FOEI for individual schools.

If the OLS and WLS estimates differ appreciably, this may be an indication that the school size, as reflected by n_g, is playing a role in the school-level achievement measure. This possibility can be easily checked by putting this term directly in the systematic part of the regression model. If an effect of size is found, it may partly be acting as a proxy for other potential school-level variables. The substantive and policy implications of including such a term in the FOEI need to be carefully considered. More generally there may be a number of school-level variables that have an effect on the school-level performance measure. The impact of school size is of particular interest because of its potential effect on the residual variance, but other school-level variables may also have explanatory power. We will denote such variables in general as $\mathbf{z}_g^{SC}$, which could be added to the systematic part of the regression model in (1).

Because the dependent variable in the regression analysis is the mean of student achievement scores it should be expected that the variance of the school achievement score around the fitted regression function would increase as the number of students contributing to the calculation of the average gets smaller. Plots of residuals obtained from a regression analyses undertaken by NSW DEC confirm this feature, particularly for schools for which the average achievement score is based on 20 or fewer students as shown in Appendix C. Use of simple weighted least squares (WLS) regression was considered to account for this feature, with the weight proportional to the number of students contributing the school average. The optimal weighting depends on the ratio of the school-level variance to the student-level variance.

As mentioned above the theoretical optimal weighting of schools in terms of the effect on the estimated regression coefficient depends on the school level and student level variance component in the multilevel model (3). From (6) we can see that the key parameter affecting the optimal weighting is ρ. Results from multilevel modelling suggest a value of ρ of about 0.06. At this value of ρ using WLS with weighting by n_g overcompensates for the variance of the residual increasing as n_g gets smaller. For example, it would lead to a school with 200 students receiving a weight 10 times that of a school with 20, whereas the school-level and student-level variance components obtained from a multilevel modelling suggests a ratio of about 1.7. Thus the equal weighting of each school implicit in OLS is closer to the optimal weighting than WLS using n_g. There are also substantive reasons to prefer OLS as each school is equally important and the FOEI should be equally applicable to all schools. The issue of large residual variation for smaller schools can be

tackled through careful treatment of outliers or robust regression as discussed in section 2.3.

The variance on the FOEI score for a school is

$$V(\tilde{y}_g \mid \overline{\mathbf{x}}_g) = \overline{\mathbf{x}}_g^T V(\hat{\beta}) \overline{\mathbf{x}}_g \tag{7}$$

Estimates of $V(\tilde{y}_g \mid \overline{\mathbf{x}}_g)$ can then be used to construct 95% confidence intervals for the FOEI score. The confidence intervals should have the usual feature of being wider at the lower and higher ends of the range of explanatory variables.

The variance of the fitted value is different from the variance of the observed residual $r_g = \overline{y}_g - \tilde{y}_g$ which is

$$V(\overline{y}_g - \tilde{y}_g) \approx V(\overline{y}_g) + V(\tilde{y}_g) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n_g} + \overline{\mathbf{x}}_g^T V(\hat{\beta}) \overline{\mathbf{x}}_g \tag{8}$$

The latter is relevant in interpreting the average performance measure for a particular school. However this is not the objective of the FOEI.

Common diagnostics involve the studentised residual, which should be calculated as

$$sr_g = \frac{r_g}{\sqrt{\hat{V}(r_g)}} \tag{9}$$

Estimation of $V(\hat{\beta})$ raises some subtle issues. If OLS or WLS is used then the standard variance estimator will assume the variance structure is σ_u^2 or $\sigma_\varepsilon^2 / n_g$ respectively. It is possible to use OLS even if we think that the variance structure follows (5) and then use so called “Huber- White” (HW) variance estimator, which is robust to the form of the residual variance.

2.3 Use of Robust Regression Methods

In any regression analysis it is important to examine the results and associated diagnostics for outliers, influential and high leverage points and multicollinearity. It is expected that there will be more variation around the fitted regression function for small schools as suggested by (5). Hence naive examination of residuals from an OLS analysis will identify many small schools as outliers, when they are consistent with the variance structure. An option is to use the studentised residual (9) where $V(\hat{\beta})$, σ_u^2 and σ_ε^2 have been estimated using multilevel modelling or using the HW variance estimator for the variance of the regression coefficients.

More formal use of robust regression methods can be considered. Such methods involve a reduction the weight given to observations with relatively large estimated residuals. This approach itself can be sensitive to the estimation of the scale factors and led to MM-estimation. This approach is available in R, which includes the option of specifying prior weight, which in

our case would be the school sample size, n_g. We can also specify the type of weight, which is essentially a frequency weight or the inverse variance weight, which is relevant for our analyses.

Because of the time involved in deciding the precise approach to and implementing multiple imputation (MI), it is easier to try different approaches with the complete case data rather than the MI data. The key feature of different approaches to the regression analysis will be clear from such an analysis. To obtain a base to which the result of the MI can be compared and to help in familiarisation of the model a large range of regression analyses based on the complete data could be carried out. The full range of analyses is:

- OLS, with and without outliers deleted, and H-W variance estimation
- WLS, with and without outliers deleted
- Multilevel model of student level data
- Robust regression, unweighted
- Robust regression, weighted (does not seem to be available in STATA)

Analysis of residuals for OLS should be based on sr_g given by (9), where the results of the multilevel model can be used to give an idea of the variance components. The residuals from the WLS will effectively use sr_g with $\sigma_u^2 = 0$. Besides the usual methods of evaluating regression models, the substantive difference between the FOEI score obtained from different analyses can be compared overall and at the school level.

The analysis focussed on the following options:

- OLS, without outliers deleted (OLS1)
- OLS, with outliers deleted (OLS2)
- WLS, without outliers deleted (WLS3)
- WLS, with outliers deleted (WLS4)
- Robust regression, unweighted (RR)

The analyses were carried out by NSW DEC using 2012 data. The analysis of RR included selective schools while the other methods did not.

Outliers were defined as those schools for which the absolute value of the studentised residual exceeded 2. The evaluation of H-W variance estimation was not seen as a priority as there is not a focus on inference about the regression coefficients and the MI approach will give variance estimates that account for the imputation process. Multilevel modelling is useful in informing us about variance comments, as we saw in section 2.2, but it is not a school-level approach, which is the currently preferred approach. As use of weights is not strongly justified in WLS it was not considered in robust regression.

The Robust Regression (RR) implemented in Stata includes an initial step that removes high-leverage outliers (based on Cook's D) and then uses an M-estimator (Huber followed by bisquare) to estimate weights for observations based on the size of the residuals.

In the robust regression 52 schools have a zero-weight, of which 31 had a NAPLAN count of less than 20. Only one of the 21 schools with zero weight and NAPLAN count greater than 20 is a comprehensive school; all others are fully selective schools.

Of the top 100 schools that have the greatest weights, 7% of them are small schools (i.e., schools with less than 20 NAPLAN students).

As expected, those observations with large studentised residual (SRED) (roughly greater than 3) have been given a weight of zero. Weights then increase as the size of the SRED decreases. Many small schools fit the regression model reasonably well, with weights from robust regression for these schools ranging from 0.8 to 1.0.

The analysis undertaken by NSW DEC showed the following features:

- While WLS (weighted by NAPLAN cohort size) and OLS1 models have produced negative (hence counter-intuitive) coefficients for the highest occupation category variable 'percentage of parents in senior management', the RR model has corrected the direction of the coefficient. Presumably this is because the effects of the influential data points have been weighted down during the robust regression estimation process.
- Of the three types of models, generally speaking, coefficients from OLS and RR models are more similar to each other than they are to those from WLS models.
- Predicted values from RR are extremely similar to those produced from OLS2 model.
- Coefficients from WLS models are more difficult to justify for use in the final FOEI model. For example, the negative coefficients associated with % of parents not in a paid work are much greater in size from the WLS models than from the OLS or RR models. In addition, while WLS models have estimated a relatively large coefficient for the percentage of parents achieving a Year 10 education level, the equivalent coefficients from OLS models and RR models are small and negative, which make sense.
- On the whole, coefficients from RR and OLS models make more sense than those from WLS. This is probably because the weighting by NAPLAN cohort size in the WLS models essentially means that all small schools are given a smaller weight than large schools. However, the OLS residual analysis demonstrates that, while most of the outliers (those with studentised residuals more than 2) are small schools, not all small schools are outliers. In fact, the majority of small schools (355 out of 439) are not outliers. They fit the model pretty well. Around 7% of the top 100 schools with the greatest weights from the robust regression modelling are small schools.
- By imposing a smaller weight for **all** small schools through WLS or excluding them, the influence of the majority of small schools, which do fit the regression model quite well has been unjustifiably reduced.

In summary, comparisons of ordinary least squares (OLS), WLS and robust regression models show that while most outliers are small schools, the

majority of small schools fit the regression model reasonably well. Robust regression reduces the weight for the outliers, that is, schools that have large residuals. Using robust regression appears to adequately account for the small schools with large residuals, while also allowing most small schools to contribute to the estimation of the regression coefficients. An alternative is an OLS model excluding selective schools and outliers for the final FOEI model. An important feature of robust regression is that it enables the use of data for schools with enrolment less than 100, which were previously excluded from the regression analyses.

An issue is whether or not to exclude selective schools from the regression estimation because of their special nature. A FOEI score can still be calculated for these schools using the estimated regression function. As the analysis carried out shows, if such schools are included in a robust regression their residuals lead to them being given zero weight anyway, hence they do not need to be manually excluded.

The issue of including community variables in the regression model and the different approaches to incorporating data for two parents are considered in sections 2.6 and 2.5 respectively.

In addition to the estimation of the variance of the FOEI score discussed in section 2.2, the results can be analysed by comparison with the FOEI that would be obtained by:

- Using OLS on the data obtained from the observed cases;
- Using OLS based on the observed + imputed data obtained using imputation
- Robust regression based on the observed cases.

This comparison should give a picture of the impact of the imputation and use of robust regression on the FOEI scores. Results of such analyses will be published through the DEC FOEI technical report.

Further research investigating the advantages and disadvantages of an explicit multilevel modelling approach using models of the form of (3) can also be undertaken for future years.

2.4 Use of School Achievement Data

The final regression analysis will use a dependent variable based on observed 2012 student achievement data for students in Year 4, 6, 8 and 10 in 2013, while the explanatory variables will be based on parental background variables for all students attending the school in 2013. The regression equation is designed to reflect the relationship between achievement in the most recent year for which achievement data are available and the parental background of students attending the school in current year. Since calculation of the FOEI score for a school is based on the parental background of all students, estimation of the regression function also uses these as the explanatory variables.

Ideally student achievement data would be available for all students in the school. One way to consider this issue is to consider the mean that would

have been obtained had students in all years been tested, $\overline{y}_g^{all}$. Then we can regard the actual means available as an estimate of this mean, with an associated measurement error, so that $\overline{y}_g = \overline{y}_g^{all} + \eta_g$. The theoretical implications of this measurement error formulation could be considered in future research.

An indirect indication of the impact of the dependent and explanatory variables being based on different sets of students could be obtained by estimating the regression coefficient using explanatory variables calculated just for the student cohorts for which achievement data are available. Using the resulting estimated regression coefficients the fitted value for a school would be calculated using the explanatory variables calculated for all students or only those for which achievement data are available, and then compared.

An option considered was using achievement scores for previous years for the dependent variable. This was not used since it would lead to the FOEI not reflecting the latest achievement data. Another option considered was the inclusion of imputed achievement data in the dependent variable, for students where achievement data was not available. It was considered that such inclusion, though would have led to most achievement data being imputed, and would generate a degree of circularity, since the imputed achievement data is based on the explanatory variables that would subsequently be used in the regression analysis. This approach would tend to reproduce the relationship between achievement score and the parent variables for the years in which the former are available. For this reason, the dependent variable in the regression model only includes the observed latest years' achievement data for each school.

2.5 Students with one parent and students with two parents

Some students have data for one parent and some have data for two parents in the data files. If there is no information for a second parent in the enrolment system, it may be assumed that the student is from a one-parent family. In the sample data file provided these students have null values for the second parent variables (as distinct from students with information on two parents in the enrolment system but where the education and occupation variables for the second parent are all coded as 'not stated').

Several approaches are available in the school-level regression modelling to deal with the parental contribution from students of different family types.

I. Calculate the school means over all parents.

II. Calculate the school means over all parents, giving each of the parents in a two-parent case a weight of 0.5.

III. Only use parent 1 from a two-parent case.

IV. Calculate the means for the one-parent variables separately to the means for the parent 1 and parent 2 variables for the two-parent cases, and put each into the regression model.

Remember that the school means are effectively proportions in each category of the parent variables.

Option I gives more weight to the parent variables for students with two parents, whereas option II effectively weights each student equally. Option III discards potentially useful information. Option IV is theoretically the best as it makes full use of the parent data, but triples the number of explanatory variables, which can affect the stability of the estimated regression coefficients. Comparison of options II and IV is an area for future research.

It was decided that in order for each student's family to count equally towards the school FOEI score, the calculation of the school-level variables assigns a weight of 2 to single parents and a weight of 1 to each parent in a two-parent family. Thus in the regression estimation and in the production of the FOEI score the school-level parental variables effectively average the characteristics of the parents in a two-parent case, which is option II.

2.6. Use of Community Variables and Other Variables in the Regression Model

The FOEI is specifically designed to account for the available data on the parents of students in the school. Other variables could be included in the regression model but would change the substantive interpretation of the FOEI and therefore not pursued here. Use of additional variables in the imputation process is considered in section 3, since a general principle is to use as many variables as reasonable in the imputation modelling, including all variables to be used in regression analysis.

The potential for school-level variables to be included in the regression modelling was briefly mentioned in section 2. It is also possible to use community level variables. For each student the value of up to 16 variables can be determined for the SA1 in which they live. School-level means for these variables can be obtained by averaging over all the students in the school, to produce measures of the community in which students live. We will denote these variables as $\mathbf{z}_g^{COM}$. Because address is an important variable there is very little missing data in the data used to calculate the community variables. The obvious way to use the community variables is to expand the regression model to include them.

There are substantive considerations involved with including such variables in the regression modelling in the calculation of the FOEI.

Another suggestion that has been made is to undertake two separate regressions, i.e. using $\overline{\mathbf{x}}_g$ and one using $\mathbf{z}_g^{COM}$ and then use the fitted value that is closest to the observed value. The statistical properties of this approach are not clear.

The regression analysis used only the parental background variables and not the community variables. Previous regression analysis performed by the DEC suggested no appreciable additional predictive power is added if the community level variables are used in the school-level regression analysis. However, community level variables are useful information to impute for missing parental data.

Indicators of ATSI status and school remoteness could also be included in the regression model, but were not included for substantive reasons. Again such

variables can be used in the imputation model.

3. Handling Missing Parent Data

Data on parents can be missing (in the data file, missing data is coded as “Not stated”). This missing data affects the estimation of the regression function and the calculation of FOEI scores for individual schools. A major issue is that there is an appreciable amount of missing parent data, which affects the calculation of $\overline{\mathbf{x}}_g$. The parent level data are used in the estimation of the regression coefficients and in the calculation of the FOEI score.

3.1 Approaches for Handling Missing Data

Missing data on the three parental background variables: highest level of school education, highest non-school education and occupation group, may affect the estimation of the regression coefficients and the calculation of the FOEI for a school.

Several approaches could be considered.

1. Estimate the regression coefficients and calculate the FOEI using only complete cases for which all explanatory variables are available.
2. Estimate the regression coefficients and calculate the FOEI using only complete cases for each explanatory variable separately.
3. Estimate the regression coefficients using only complete cases for which all explanatory variables are available but calculate the FOEI including values imputed for the missing values.
4. Estimate the regression coefficients and calculate the FOEI including values imputed for the missing values.

Approach (1) leads to an appreciable reduction in the number of students on which school means are based and those that are used may lead to biased values. Approach (2) reduces these problems somewhat, but the basic problems remain and the means for different variables are based on different sets of students. Approach (3) has some value, but leads to the regression model using explanatory variables with different values in the estimation of the regression coefficients and the calculation of the FOEI. Approach (4) maximises the use of the observed data and uses the same values of the school means in the estimation of the regression coefficients and the calculation of the FOEI, therefore is used in the calculation of 2013 FOEI.

Imputation can be implemented using multiple imputation (MI) based on a chained equation approach. For a student each missing variable will be imputed using all the other observed variables, including the response variable, the other explanatory variables to be used in the regression model, plus several additional variables. The imputation is applied at the student level and then the resulting imputed and observed data, which is called the *completed data* is used to calculate school-level means that are then used in the estimation of the regression coefficients.

Multiple imputation, its related assumptions and benefits, are described in the following section. Issues affecting the imputation in this context will be considered in subsequent sections.

3.2 Multiple Imputation (MI)

A model-based multiple imputation approach has been evaluated using multiple imputations by chained equations (MICE). This approach uses statistical models that reflect the relationships between the variables in the observed data to impute plausible values for the missing data. These statistical models are referred to as the *imputation model*. The imputation is carried out multiple times (M=10) to enable valid estimates of uncertainty accounting for both the regression model estimation and the imputation itself. This is a widely adopted and flexible approach that allows the full use of the observed data for many variables that can be implemented using readily available statistical software (White et al, 2011).

As with any imputation approach the method of MI is based on some assumptions, including the assumption that conditional on the observed data the unobserved data are missing at random. This missing at random (MAR) assumption is less restrictive than the assumption of Missing Completely at Random (MCAR) which is assumed if complete cases, that is students for whom all parental data are available, only are used in the analysis. It is possible that the mechanism of missingness is Missing Not at Random (MNAR), however, the use of a large number of explanatory variables in the imputation model should assist in moving closer to the MAR assumption and therefore reduce any possible bias (White et al, 2011).

A major benefit of using MI is that it enables the variance contributed by the imputation process to be included in the estimation of the variance of the estimated parameter. For a parameter θ, let $\hat{\theta}_m$ be the estimated value calculated from the *m*th completed data set. Then by applying Rubin's rules (Rubin, 1987)

$$\hat{\theta}^{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{10}$$

is the MI estimator and its variance is estimated by

$$V(\hat{\theta}^{MI}) = \hat{V} + \frac{M+1}{M}\hat{B} \tag{11}$$

where $\hat{V} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}(\hat{\theta}_m)$ and $\hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}^{MI}\right)^2$.

This approach can be applied to the estimation of the regression parameters, so $\theta = \beta$, and the FOEI score, so $\theta = \bar{\mathbf{x}}_g^T\beta$. An overview of the process is shown in Figure 1.

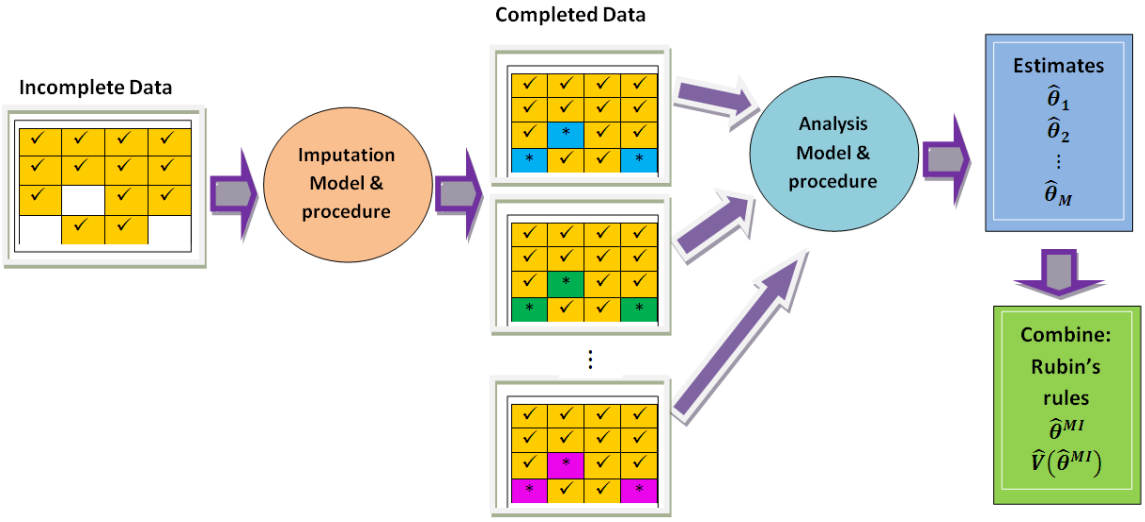

Figure 1: Overview of multiple imputation and resulting analysis using Rubin's rules.

Application of MI to the FOEI score required that the imputation is applied simultaneously to the regression coefficient and the school means of the explanatory variables. This means that

$$\tilde{y}_g^{MI} = \frac{1}{M}\sum_{m=1}^{M}\tilde{y}_{gm} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\mathbf{x}}}_{gm}^{T}\hat{\beta}_m \qquad (12)$$

This value is not the same as applying the MI estimate of the regression coefficient to the MI estimate of the school means, i.e. $\tilde{y}_g^{MI(1)} = \hat{\bar{\mathbf{x}}}_g^{MI^T}\hat{\beta}^{MI}$. It is also not necessarily the same as the fitted value that would be obtained from regression using the MI estimate of school means, $\hat{\bar{\mathbf{x}}}_g^{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\mathbf{x}}}_{gm}$, as the explanatory variables, which we can denote by $\tilde{y}_g^{MI(2)}$. For discussion with school principals there will naturally be interest in the imputed values of the school means of the explanatory variables, which is just $\hat{\bar{\mathbf{x}}}_g^{MI}$. For this reason, NSW DEC used $\tilde{y}_g^{MI(2)}$, which is based on using $\hat{\bar{\mathbf{x}}}_g^{MI}$ as independent variables, in the FOEI calculation process, as it allows the department to better explain the school level means of the explanatory variables and the associated weights that are used to calculate FOEI to the principals. Further work remains to examine the practical differences in the fitted values, $\tilde{y}_g^{MI}$, produced from equation (12) and $\tilde{y}_g^{MI(2)}$.

3.3 One and two-parent families

As explained in Section 2.5, some students have data for one parent and some have data for two parents in the data files. Students with null values for the second parent variables (reflecting the absence of a second parent in the enrolment system) may be assumed to be from a one-parent family. These students are distinct from students with information on two parents in the

enrolment system but where the education and occupation variables for the second parent are all coded as 'not stated' in data files provided.

It was considered important to distinguish between one and two-parent families in the imputation model so as not to impute for a second parent in a one-parent family and to also allow for the use of information on the second parent in a two-parent family. The imputation for one-parent and two-parent data was therefore carried out separately, on the basis that the relationship between the parent background variables is different for the one parent variables and the equivalent variables in the two-parent data file, as was demonstrated by the department's own analysis.

The data file was separated into two data sets, based on a set of null values for all of the second parent variables, and imputation carried out separately. There is a sufficient number of students from both one and two-parent families to enable the relationships amongst parent variables to be estimated separately for the two groups of students.

For the purpose of the review, any student who had missing ('not stated') values on all six of the parent background variables (around 5% of the students in the sample data set) was deleted from the data used for the imputation.

3.4 The multiple imputation model

MI is based on estimating the relationship between variables using cases (i.e. students) for which the variables are available. The set of variables in the imputation model should be larger and broader in scope than that required for the ultimate analytic model that will be applied to the completed data set (Heeringa et al, 2010, section 11.4). So the imputation model is a key element of MI and should include all the covariates that may be in the analysis model and in particular must contain the analysis model outcome variable (Moons et al, 2006).

For the purpose of this review, a sample of 2012 data was provided by NSW DEC to NIASRA. The sample data included approximately 50% of schools with all students at the selected schools. Data included student year of schooling, selective school category, gender, Aboriginal status, reported and standardised NAPLAN reading and numeracy results, parent education and occupation variables, and a series of community variables derived from the 2011 ABS census. The schools are categorised as fully selective, partially selective and non-selective.

In the imputation process, plausible values for the missing data were imputed for the three categorical parental background variables, for which details are given below. For students belonging to single parent families, there are three parental background variables included in the MI model. For the two-parent cases, there are two of each of the variables described below: one set for parent 1 and another set for parent 2. Hence, for the students belonging to two-parent families, the imputation was carried out for each of the six variables.

- Parent highest education level
    - Year 9 or equivalent or below;

    - Year 10 or equivalent;
    - Year 11 or equivalent;
    - Year 12 or equivalent.
- Parent highest non-school education level;
    - No non-school education;
    - Certificate I to IV (including trade certificate);
    - Advanced diploma or diploma;
    - Bachelor degree or above.
- Parent occupation group
    - Senior management;
    - Other business manager, arts/media/sport;
    - Tradesperson, clerks, sales and service;
    - Machine operators, hospitality staff etc.;
    - Not in paid work.

The model includes the parental background variables along with the combined standardised student achievement scores for numeracy and literacy, ATSI status, school remoteness and a set of ten community variables derived from the 2011 ABS census in the same statistical area (SA1) as the student's address. The imputation excluded students from fully selective schools.

The original variable for the ATSI status of the student had five categories:

- ATSI status (original)
    - Aboriginal but not Torres Strait Islander Origin
    - Torres Strait Islander but not Aboriginal Origin
    - Both Aboriginal and Torres Strait Islander Origin
    - Neither Aboriginal nor Torres Strait Islander Origin
    - Not stated/Unknown

The first three categories were collapsed into one category to form the revised ATSI variable used in the imputation models:

- ATSI status (revised)
    - Aboriginal and/or Torres Strait Islander Origin
    - Neither Aboriginal nor Torres Strait Islander Origin
    - Not stated/Unknown

The remoteness of the school using the MCEECDYA remoteness classification with the following three categories:

- Remote_code
    - Metropolitan
    - Provincial

- Remote (including Very Remote)

The ten census variables were chosen based on previous analysis carried out by DEC, and consist of the following variables:

1) Percentage of people with annual household income between $10400 and $20799 (INC_BET 10400_20799).
2) Percentage of people 15 years and over with advanced diploma or diploma qualifications (TE_DIPLOMA).
3) Percentage of people 15 years and over with no post-school qualifications (TE_NONE).
4) Percentage of people 15 years and over whose highest level of schooling completed is Year 11 or lower (SE_Y11_LOWER).
5) Percentage of people in the labour force who are unemployed (LABOUR_UNEMP).
6) Percentage of employed people who work in a skill level 1 occupation (high) (OCC_GRP1_HIGH).
7) Percentage of employed people who work in a skill level 4 occupation (mid) (OCC_GRP4_MID).
8) Percentage of employed people who work in a skill level 5 occupation (low) (OCC_GRP5_LOW).
9) Percentage of occupied dwellings with no internet connection (INTERNET_NONE).
10) Percentage of families that are one-parent families with dependent offspring only (SINGLE_P_VS_FAMILY_C).

The choice of the census variables to include in the imputation model is an issue that could be considered for further research in this context.

For approximately 5% of cases, an SA1 could not be matched to student address and therefore items on the community variables are missing for some students. Thus, the community variables used as explanatory or independent variables in the imputation model may contain missing values. In Stata, the imputation model can proceed (if the `force' option is specified), however, missing data on the parent variables will not be imputed for these cases and a listwise deletion approach is adopted.

It is recommended that quality checking of student addresses and subsequent matching to SA1 geographic codes be carried out in preparation for the imputation in order to minimize any missing items in the explanatory variables.

Consideration was also given to the inclusion of additional variables in the imputation model. These additional variables included the students' grade (or year level) and sex. As there was no practical or theoretical basis on which to suggest that students' grade and sex are likely to be predictive of parental background information, or the relationships between parental background information and the other variables, these additional variables were not included in the imputation model.

An ordinal logistic model is applied for the imputation of the two ordinal variables: *Parent highest education level* and *Parent highest non-school education level*. As the *Parent occupation group* is considered as a nominal

variable rather than an ordinal variable, a multinomial logistic model is applied for the imputation of the missing values for this variable. No scoring of categories was used.

To evaluate different multiple imputation models, the multiple imputation for the parent background variables was carried out for the review sample data using three different models, which are described below. Table 1 and Table 2 set out the details for the imputation models used for the single and two-parent cases respectively. The dependent variable is the variable for which item-missing data will be imputed. The explanatory or predictor variables are variables that will improve the precision and accuracy of the imputation of item-missing data (Heeringa et al, 2010, section 11.4.1) such as ATSI status and remoteness. These should also include any variables to be used in the analytic (i.e. regression) model, such as the student mean of achievement scores, and the parent background variables. Other variables which are likely to predict the propensity for response, such as census variables in this context, will help to reduce any bias associated with the assumed MAR data mechanism (Heeringa et al, 2010, p352).

As can be seen from Tables 1 and 2, the three models include different sets of explanatory variables. Model 1 includes all the parent variables and student means of achievement scores, model 2 adds the 10 community variable obtained from the census and model 3 adds ATSI status and an indicator of the remoteness of the school. Model 1 is the simplest to Model 3 having the most explanatory variables.

Table 1: Single parent cases: details of the imputation models for the parent background variables.

Dependent variable	Model	Explanatory variables			
PG1_ School_Educ	Model 1	PG1_NonSchool_Educ PG1_Occ_Group	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness
PG1-NonSchool_Educ	Model 1	PG1_School_Educ PG1_Occ_Group	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness
PG1_Occ_Group	Model 1	PG1_School_Educ PG1_NonSchool_Educ	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness

Table 2: Two-parent cases: details of the imputation models for the parent background variables.

Dependent variable	Model	Explanatory variables			
1st Parent					
PG1_ School_Educ	Model 1	PG1_NonSchool_Educ PG1_Occ_Group PG2_School_Educ PG2_NonSchool_Educ PG2_Occ_Group	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness
PG1-NonSchool_Educ	Model 1	PG1_School_Educ PG1_Occ_Group PG2_School_Educ PG2_NonSchool_Educ PG2_Occ_Group	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness
PG1_Occ_Group	Model 1	PG1_School_Educ PG1_NonSchool_Educ PG2_School_Educ PG2_NonSchool_Educ PG2_Occ_Group	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness
2nd Parent					
PG2_ School_Educ	Model 1	PG1_School_Educ PG1_NonSchool_Educ PG1_Occ_Group PG2_NonSchool_Educ PG2_Occ_Group	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness
PG2-NonSchool_Educ	Model 1	PG1_School_Educ PG1_NonSchool_Educ PG1_Occ_Group PG2_School_Educ PG2_Occ_Group	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness
PG2_Occ_Group	Model 1	PG1_School_Educ PG1_NonSchool_Educ PG1_Occ_Group PG2_School_Educ PG2_NonSchool_Educ	StudMean		
	Model 2			10 Census variables	
	Model 3			10 Census variables	ATSI Remoteness

3.5 Schools as clusters

The imputation does not explicitly use the school indicator variable, although the imputation uses the community variables and therefore reflects some characteristics of the local community. Imputation taking into account the nesting structure of the data (i.e., students are clustered in schools) was not adopted because of concerns about the stability of relationships based on small numbers of responding cases in many schools. Some of the larger schools would have had sufficient cases to consider imputation within the school, but then the methods of imputation would have differed across schools, which was considered unreliable.

The imputation can be done within strata, which effectively means putting strata and the interaction of strata within each explanatory variable in the model. An option would be to put only main effects in the imputation model. Each school could be considered as a stratum, which allows the relationship between the variable to vary between schools. It also means the relations will be estimated using the data for each school separately.

The multilevel structure of the data is an issue to be considered for further research concerning imputation in this context.

3.6 Availability of school achievement data for students for several years

The standardised score used in the imputation for a student is the average of the standardised NAPLAN reading and numeracy scores if both are available; otherwise if only one is available, that score is used. Standardising is based on the NSW government means and standard deviations and is performed separately for each grade or year level cohort. For a particular calendar year student achievement data for NAPLAN are not available for all students in all years; only students in Years 3, 5, 7 and 9 are tested in any given calendar year.

There are three approaches to dealing with this issue in the imputation process.

a. Use the latest year's achievement data in the imputation models (eg use 2012 achievement data for the imputation of parental data in 2013).

b. Expand option (a) to include achievement data from the previous years to increase the number of students whose missing parental information can be imputed.

c. Use option (b), and to further increase the number of students whose missing parental information can be imputed, impute the achievement data for the cohorts for which achievement data are not available.

If Option (a) were adopted, only two-sevenths of the students in a primary school and one-third of the students in a secondary school would be included in the imputation modelling. As a result, missing parent data would not be imputed for the large majority of students. This would have a flow-on effect on the regression modelling as explanatory variables would contain many missing items.

Option (b) includes achievement data from the previous years in order to increase the number (or proportion) of all students who have an achievement score that can be used in the imputation model. To maximise the number of students for which standardised achievement data are available for use in imputing the parent variables, standardised NAPLAN reading and numeracy results for students enrolled in a school in 2013 obtained from 2010, 2011 and 2012 can be used. This allowed standardised achievement data to be used for Years 4 -12 for students in 2013. Table 3 illustrates this concept.

Table 3: Combining Student achievement scores across calendar years

Year in 2013	NAPLAN Data Used in Imputation
Kindergarten to YEAR 3 (as at April in 2013)	None available
YEAR 4	2012 YEAR 3
YEAR 5	2011 YEAR 3
YEAR 6	2012 YEAR 5
YEAR 7	2011 YEAR 5
YEAR 8	2012 YEAR 7
YEAR 9	2011 YEAR 7
YEAR 10	2012 YEAR 9
YEAR 11	2011 YEAR 9
YEAR 12	2010 YEAR 9

Although the 2010 NAPLAN data was not provided in the sample data set for the purpose of this review, it could be included in the actual analysis for the entire data, and hence results for matched students in Years 4-12 in 2013 could be utilised.

In option (c), an achievement score could be missing in two ways: firstly, it could be missing (by design) if a student is not in a targeted cohort for 2010, 2011 or 2012 (eg Kindergarten to Year 3 students in 2013). Secondly, student achievement data could also be missing for students in Years 3 to 10 for a given calendar year if the student was either absent, withdrawn or exempted from sitting the test.

Where an achievement score is missing, it could be included in the imputation model as one of the variables to be imputed. For the students who were withdrawn or absent from sitting the test (approximately 3.8% of targeted cohorts in the sample data), it is plausible to impute an achievement score based on the other variables in the imputation model, including the parent background variables.

However there are other considerations regarding imputing missing parental data for students who have missing achievement data by design or due to exemption. These are covered in the next section.

3.7 Imputation of parents' data for students without achievement scores (due to exemption or by design)

A particular group of students that we need to consider are those that are exempted for NAPLAN and therefore do not have any NAPLAN based achievement scores. Students can be exempted for reasons such as intellectual disability or limited English proficiency. Another group that needs consideration are those missing achievement data by design (eg Kindergarten to Year 3 as at April in 2013). It is intended that the parental data for such students be included when calculating the FOEI score for a school. Hence the issue is how to treat these students when some of the parental data are missing.

There appears to be four options:

1) Drop the cases from the imputation process and from the regression model.
2) Include these cases with all other students in the imputation process that includes the achievement data as a data item to be imputed when it is not available.
3) Include these cases with all other students in an imputation process that does not include achievement data as an explanatory variable or a variable to be imputed.
4) Combine all cases with no achievement data as a separate group and run an imputation model that does not include achievement data, either as an explanatory variable or a variable to be imputed.

Note these imputation options are considered in the context that they are used to generate imputed parental values for those students with no achievement data either by design or by exemption.

Consider the assumptions and issues in each option:

Option (1) effectively imputes the school average of the parental data for the deleted cases.

Option (2) involves imputing parental data as well as achievement scores using all explanatory variables including achievement scores for all students.

Option (3) involves imputing parental values for all students using explanatory variables other than the achievement variable.

Option (4) uses the same imputation model as option (3) but limits the imputation process to only those students with no achievement scores, either by design or by exemption.

Options (1) and (4) seem inferior to options (2) and (3). It needs to be remembered that the imputation process does not assume that the set of parents that are treated together have the same characteristics. It assumes that the relationship between observed and missing variables is the same. While the imputed achievement score under option (2) would not be what the students would have obtained had they sat the NAPLAN test, this is not relevant. The question is whether there are any reasons to think that the relationship between missing and observed parent data is any different for the parents of students with no achievement scores as the relationship for all parents. If not then option (2) should be preferred as it treats all parents in the same way. For cases when achievement scores are not available the imputed values of parental data are based on the relationship of that variable with the other parental variables and the imputed achievement score. As the achievement score itself is imputed based on the relationship of the achievement score and the observed parental variables, the approach is effectively using the observed parent variables. This means that option 2 and option 3 are unlikely to produce significantly different imputed parental values for students with no achievement data.

From discussion with the department, imputation of achievement data for students where it is missing either by design or by exemption is unlikely to be accepted by stakeholders. This could ultimately affect the acceptance of

FOEI values by school community. On this basis option (3) was the option finally accepted into the business rules by DEC.

3.8 Imputation Process

Given the discussion in section 3.3 to 3.7, Figure 2 provides a high-level description of different subsets of student cohorts (based on the 2012 sample data) for which different imputation models might need to be applied.

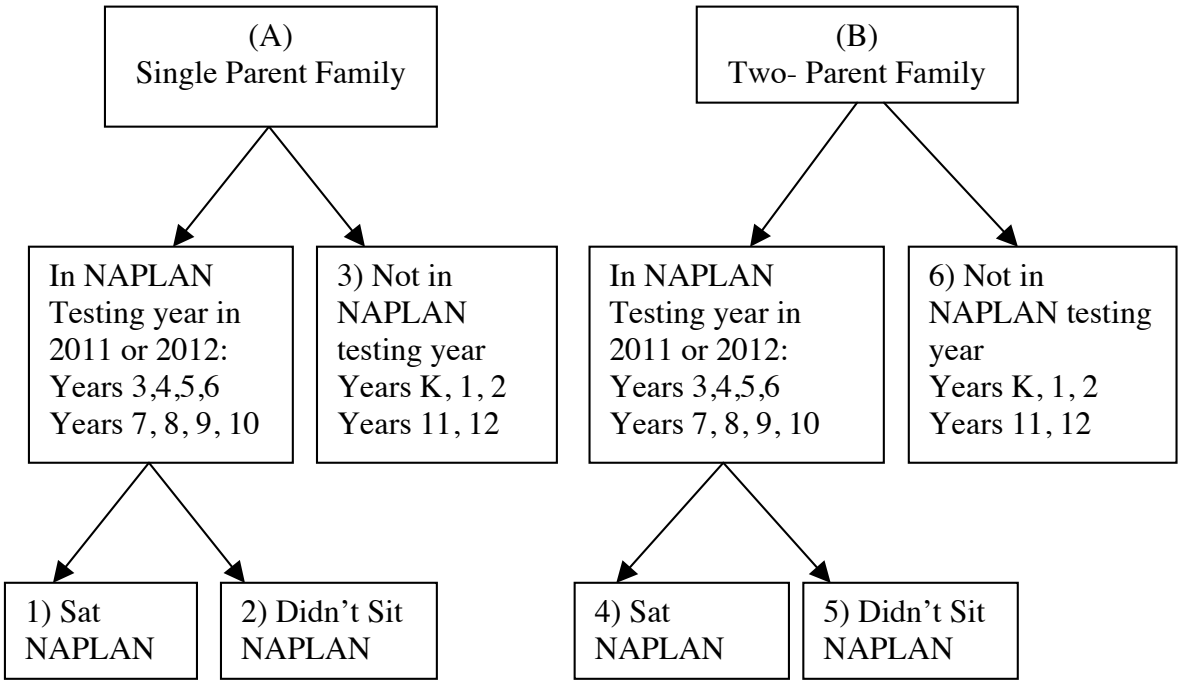

Figure 2: High-level conceptualisation of the different sets of student cohorts for which different imputation models might need to be applied (based on the 2012 sample data).

3.9 Results of Multiple Imputation Review

The results of the multiple imputation models explored in the review are summarised in Appendix A for the single parent cases and in Appendix B for the two parent cases. The results are for each of the three models defined in Tables 1 and 2 for which NIASRA ran the imputation process on the review data set. Note that all models use achievement data as an explanatory variable (not an item for imputation) and are only run for students with achievement data.

As part of the imputation process, as shown in Figure 1, new variables are created which include the imputed items as well as the original observed items. As the number of imputations was set to M=10, ten new variables are created for each parent background variable. In Stata 12, diagnostic tables are available (using the *midiagplots* function) for the completed categorical variables. A table showing the proportion of each category obtained under Observed, Imputed and Completed data is created for each value of m=1, …, 10. As a consequence, ten tables are produced for each variable imputed.

This results in 30 tables for the single parent cases and 60 tables for the two-parent cases for each of the three different models that were run.

To summarise the resulting data, the mean proportion and standard deviation were calculated for each set of 10 tables for each category of the variable. For example, for the variable *PG1_School_Educ*, the proportion of students from a single parent family whose parent's highest school education level attained was a *Year 9 or equivalent or below* was calculated from each imputed data set and then the mean and standard deviation of those values are reported here. These results do not include the imputation of any student achievement data. In the review data set there were a total of 59849 students identified as belonging to a one-parent family.

The overall change to the mean proportions when the imputed data is included is not significant. For example, for Single Parents, school education level of *Year 9 or equivalent or below* has an observed proportion of 14.4%, (refer to Appendix A1) the mean proportion over all ten imputed under Model 1 is 14.9% giving an overall mean proportion on the completed data of 14.4%. The movement for this category is as expected: that is, conditional on the relationships of other explanatory variables, the parents who did not respond to the question are more likely to have attained *Year 9 or equivalent or below* for their school education than those who did answer the question. For Models 2 and 3, the mean proportion over all ten proportions for cases with imputed data increases to 15.2% and 15.1% respectively, however the mean completed proportion only changes slightly to 14.5% for both models. This shows that the additional explanatory variables used in Models 2 and 3 have had an effect on the imputed proportions. The reason the overall average proportion based on the observed + imputed data has not changed significantly reflects the fact that only 9% of the parents were missing school education levels.

Another example of the effect of the additional explanatory variables can be seen for the imputation of missing items in the variable Occupation Group (refer to Appendix A3). The observed proportion of single parents not in paid work is 34.9%; for Model 1, the mean proportion imputed is 38.5%, and for completed it is 35.9%. The mean imputed proportion for *Not in paid work* increases to 40.1% for Model 2 and to 40.4% for Model 3. This shows that the additional explanatory variables in Model 2 and 3 in the multiple imputation procedure are having an effect on the final estimates. The direction of this change is as expected.

The proportions for each category on the completed data are similar over all ten imputations. The standard deviation reported in the results is the standard deviation across all ten proportions. Since the observed proportions do not change from model to model, the standard deviations are all zero for the observed proportions for each model. The standard deviations reported for the completed proportions are small in the order of 10^{-3} to 10^{-4}. This small standard deviation of the results of the ten imputations seems to suggest that the choice of M =10 was adequate.

3.10 Final business rules for implementation

After the review and further discussion, DEC decided to adopt the following steps for the imputation on the entire file:

- Step 1: Separate student records to single parent and two parent records and apply the following imputation processes separately on the two sets of records.
- Step 2: Select students with NAPLAN results or are in the NAPLAN cohorts but were absent or withdrawn, run an imputation model imputing for missing values in the parent background variables (PBG) (three variables for the single parent records, and six variables for the two-parent records) and student achievement mean (average standardised score from NAPLAN), using PBG, student achievement mean, ATSI, school remoteness and community variables. Save as a separate file.
- Step 3: Select all students. Run an imputation model imputing for missing values in PBG (three variables for the single parent records, and six variables for the two-parent records)), using PBG, ATSI, school remoteness and community variables (i.e. exclude student achievement mean from explanatory variables).
- Step 4: From results of Step 3, select students that are not in NAPLAN cohorts or were exempted from NAPLAN, and merge them with the results from Step 2.
- Step 5: Merge the two imputed data files generated for single and two-parent student records into a single file comprising all students.

4. Imputation Methodology and Regression Analysis: Conclusions

In section 2 the issues associated with the approach to the regression analysis were considered and in section 3 the issues associated with imputation were considered. From these considerations and the analyses conducted an approach has been developed for which the key features are:

- student level Imputation of missing parent data using a multiple imputation by chained equations approach involving the parent variables themselves, community variables, NAPLAN based student achievement scores, ATSI status and school remoteness.
- Use school level robust regression to construct the regression model.
- The dependent variable in the regression analysis is the observed student achievement scores obtained from 2012 student and NAPLAN data. These refer to students in Years 4, 6, 8 and 10 in 2013.
- In the regression analyses the explanatory variables are the parental background variables. These are calculated using parental background data for all students in the school in 2013, including the imputed values for missing data.
- All students enrolled in the school in 2013 are used in the calculation of the FOEI.
- The variance of the estimated FOEI score for a school can be estimated accounting for the imputation process.

References

Heeringa, S.G., West, B.T. and Berglund, P.A. (2010) Applied Survey Data Analysis. Chapman & Hall / CRC: Boca Raton, FL, U.S.A.

Moons, K.G.M., Donders, R. A. R.T., Stijnen, T., and Harrell, F. E. (2006) Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, Vol. 59, p1092-1101.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York.

White, I.R, Royston, P. and Wood, A.M. (2011) Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. Vol. 30, p377-399.

Appendix A: Summary of Multiple Imputation for Single Parent Families:

Appendix A 1: Results for Level of School_Education

Model 1: NonSchool_Educ + Occ_Group + StudMean

Number	of	observed	=	34555
Number	of	imputed	=	5474
Number	of	completed	=	40029

Model 1

mean	code	PG1b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.144	0.149	0.144
	2	Year 10 or equivalent	0.377	0.367	0.376
	3	Year 11 or equivalent	0.103	0.097	0.103
	4	Year 12 or equivalent	0.375	0.388	0.377
sd	1	Year 9 or equivalent or below	0.00	0.00473	0.00065
	2	Year 10 or equivalent	0.00	0.00678	0.00093
	3	Year 11 or equivalent	0.00	0.00203	0.00028
	4	Year 12 or equivalent	0.00	0.00751	0.00103

Model 2: NonSchool_Educ + Occ_Group + StudMean + 10 Census variables

Number	of	observed	=	34555
Number	of	imputed	=	5174
Number	of	completed	=	39729

Model 2

mean	code	PG1b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.144	0.152	0.145
	2	Year 10 or equivalent	0.377	0.370	0.376
	3	Year 11 or equivalent	0.103	0.097	0.103
	4	Year 12 or equivalent	0.375	0.381	0.376
sd	1	Year 9 or equivalent or below	0.00	0.00412	0.00054
	2	Year 10 or equivalent	0.00	0.00527	0.00069
	3	Year 11 or equivalent	0.00	0.00467	0.00061
	4	Year 12 or equivalent	0.00	0.00755	0.00098

Model 3: NonSchool_Educ + Occ_Group + StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	34555
Number	of	imputed	=	5148
Number	of	completed	=	39703

Model 3

mean	code	PG1b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.144	0.151	0.145
	2	Year 10 or equivalent	0.377	0.369	0.376
	3	Year 11 or equivalent	0.103	0.099	0.103
	4	Year 12 or equivalent	0.375	0.380	0.376
sd	1	Year 9 or equivalent or below	0.00	0.00280	0.00052
	2	Year 10 or equivalent	0.00	0.00461	0.00067
	3	Year 11 or equivalent	0.00	0.00494	0.00074
	4	Year 12 or equivalent	0.00	0.00624	0.00094

Appendix A 2: Results for Level of _NonSchool_Education

Model 1: School_Educ + Occ_Group + StudMean

Number	of	observed	=	27799
Number	of	imputed	=	12230
Number	of	completed	=	40029

mean	code	Model1 PG1b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.324	0.395	0.346
	5	Certificate I to IV (including trade cer	0.395	0.396	0.396
	6	Advanced diploma/Diploma	0.141	0.119	0.135
	7	Bachelor degree or above	0.139	0.090	0.124
sd	4	No non-school education	0.00	0.00499	0.00153
	5	Certificate I to IV (including trade cer	0.00	0.00349	0.00107
	6	Advanced diploma/Diploma	0.00	0.00293	0.00090
	7	Bachelor degree or above	0.00	0.00353	0.00108

Model 2: School_Educ + Occ_Group + StudMean + 10 Census variables

Number	of	observed	=	27799
Number	of	imputed	=	11506
Number	of	completed	=	39305

mean	code	Model 2 PG1b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.324	0.403	0.347
	5	Certificate I to IV (including trade cer	0.395	0.398	0.396
	6	Advanced diploma/Diploma	0.141	0.117	0.134
	7	Bachelor degree or above	0.139	0.083	0.122
sd	4	No non-school education	0.00	0.00392	0.00115
	5	Certificate I to IV (including trade cer	0.00	0.00515	0.00151
	6	Advanced diploma/Diploma	0.00	0.00280	0.00082
	7	Bachelor degree or above	0.00	0.00246	0.00072

Model 3: School_Educ + Occ_Group + StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	27799
Number	of	imputed	=	11471
Number	of	completed	=	39270

mean	code	Model 3 PG1b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.324	0.405	0.348
	5	Certificate I to IV (including trade cer	0.395	0.393	0.395
	6	Advanced diploma/Diploma	0.141	0.117	0.134
	7	Bachelor degree or above	0.139	0.085	0.123
sd	4	No non-school education	0.00	0.00404	0.00123
	5	Certificate I to IV (including trade cer	0.00	0.00379	0.00120
	6	Advanced diploma/Diploma	0.00	0.00338	0.00107
	7	Bachelor degree or above	0.00	0.00193	0.00067

Appendix A 3: Results for Occupation_Group

Model 1: School_Educ + NonSchool_Educ + StudMean

Number	of	observed	=	28622
Number	of	imputed	=	11407
Number	of	completed	=	40029

mean	code	Model 1 PG1b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.083	0.069	0.079
	2	Other business manager, arts/media/sport	0.133	0.116	0.128
	3	Tradesman/woman, clerks, sales and servi	0.216	0.202	0.212
	4	Machine operators, hospitality staff, as	0.220	0.228	0.222
	8	Not in paid work	0.349	0.385	0.359
sd	1	Senior management	0.00	0.00246	0.00070
	2	Other business manager, arts/media/sport	0.00	0.00342	0.00098
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00529	0.00151
	4	Machine operators, hospitality staff, as	0.00	0.00445	0.00127
	8	Not in paid work	0.00	0.00641	0.00183

Model 2: School_Educ + NonSchool_Educ + StudMean + 10 Census variables

Number	of	observed	=	28622
Number	of	imputed	=	10649
Number	of	completed	=	39271

mean	code	Model 2 PG1b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.083	0.065	0.078
	2	Other business manager, arts/media/sport	0.133	0.110	0.127
	3	Tradesman/woman, clerks, sales and servi	0.216	0.197	0.211
	4	Machine operators, hospitality staff, as	0.220	0.227	0.222
	8	Not in paid work	0.349	0.401	0.363
sd	1	Senior management	0.00	0.00274	0.00074
	2	Other business manager, arts/media/sport	0.00	0.00320	0.00087
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00236	0.00064
	4	Machine operators, hospitality staff, as	0.00	0.00673	0.00183
	8	Not in paid work	0.00	0.00528	0.00143

Model 3: School_Educ + NonSchool_Educ + StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	28622
Number	of	imputed	=	10615
Number	of	completed	=	39237

mean	code	Model 3 PG1b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.083	0.068	0.079
	2	Other business manager, arts/media/sport	0.133	0.108	0.126
	3	Tradesman/woman, clerks, sales and servi	0.216	0.195	0.210
	4	Machine operators, hospitality staff, as	0.349	0.404	0.364
	8	Not in paid work	0.349	0.404	0.364
sd	1	Senior management	0.00	0.00300	0.00082
	2	Other business manager, arts/media/sport	0.00	0.00227	0.00079
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00384	0.00103
	4	Machine operators, hospitality staff, as	0.00	0.00576	0.00151
	8	Not in paid work	0.00	0.00576	0.00151

Appendix B: Summary of Multiple Imputation for Two- Parent Families

Appendix B1: Results for Level of _School_Education (Parent 1)

Model 1: StudMean

Number	of	observed	=	182159
Number	of	imputed	=	29326
Number	of	completed	=	211485

mean	code	Model 1 PG1b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.077	0.085	0.078
	2	Year 10 or equivalent	0.295	0.295	0.295
	3	Year 11 or equivalent	0.083	0.079	0.082
	4	Year 12 or equivalent	0.545	0.541	0.545
sd	1	Year 9 or equivalent or below	0.00	0.00163	0.00000
	2	Year 10 or equivalent	0.00	0.00288	0.00047
	3	Year 11 or equivalent	0.00	0.00204	0.00052
	4	Year 12 or equivalent	0.00	0.00343	0.00057

Model 2: StudMean + 10 Census variables

Number	of	observed	=	182159
Number	of	imputed	=	28250
Number	of	completed	=	210409

mean	code	Model 2 PG1b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.077	0.086	0.078
	2	Year 10 or equivalent	0.295	0.297	0.295
	3	Year 11 or equivalent	0.083	0.079	0.082
	4	Year 12 or equivalent	0.545	0.538	0.544
sd	1	Year 9 or equivalent or below	0.00	0.00149	0.00000
	2	Year 10 or equivalent	0.00	0.00250	0.00042
	3	Year 11 or equivalent	0.00	0.00176	0.00048
	4	Year 12 or equivalent	0.00	0.00362	0.00063

Model 3: StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	182159
Number	of	imputed	=	28206
Number	of	completed	=	210365

mean	code	Model 3 PG1b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.077	0.086	0.078
	2	Year 10 or equivalent	0.295	0.296	0.295
	3	Year 11 or equivalent	0.083	0.079	0.082
	4	Year 12 or equivalent	0.545	0.540	0.545
sd	1	Year 9 or equivalent or below	0.00	0.00210	0.00032
	2	Year 10 or equivalent	0.00	0.00247	0.00032
	3	Year 11 or equivalent	0.00	0.00222	0.00042
	4	Year 12 or equivalent	0.00	0.00295	0.00052

Appendix B2: Results for Level of _NonSchool_Education (Parent 1)

Model 1: StudMean

Number	of	observed	=	155391
Number	of	imputed	=	56094
Number	of	completed	=	211485

mean	code	Model1 PG1b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.247	0.334	0.270
	5	Certificate I to IV (including trade cer	0.321	0.349	0.329
	6	Advanced diploma/Diploma	0.167	0.153	0.163
	7	Bachelor degree or above	0.265	0.166	0.238
sd	4	No non-school education	0.00	0.00222	0.00057
	5	Certificate I to IV (including trade cer	0.00	0.00211	0.00067
	6	Advanced diploma/Diploma	0.00	0.00117	0.00032
	7	Bachelor degree or above	0.00	0.00207	0.00070

Model 2: StudMean + 10 Census variables

Number	of	observed	=	155391
Number	of	imputed	=	54197
Number	of	completed	=	209588

mean	code	Model 2 PG1b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.247	0.338	0.271
	5	Certificate I to IV (including trade cer	0.321	0.351	0.329
	6	Advanced diploma/Diploma	0.167	0.151	0.163
	7	Bachelor degree or above	0.265	0.160	0.238
sd	4	No non-school education	0.00	0.00221	0.00067
	5	Certificate I to IV (including trade cer	0.00	0.00264	0.00082
	6	Advanced diploma/Diploma	0.00	0.00223	0.00067
	7	Bachelor degree or above	0.00	0.00088	0.00032

Model 3: StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	155391
Number	of	imputed	=	54125
Number	of	completed	=	209516

mean	code	Model 3 PG1b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.247	0.337	0.270
	5	Certificate I to IV (including trade cer	0.321	0.351	0.329
	6	Advanced diploma/Diploma	0.167	0.152	0.163
	7	Bachelor degree or above	0.265	0.160	0.238
sd	4	No non-school education	0.00	0.00231	0.00070
	5	Certificate I to IV (including trade cer	0.00	0.00286	0.00074
	6	Advanced diploma/Diploma	0.00	0.00114	0.00042
	7	Bachelor degree or above	0.00	0.00116	0.00042

Appendix B3: Results for _Occupation _Group (Parent 1)

Model 1: StudMean

Number	of	observed	=	164720
Number	of	imputed	=	46765
Number	of	completed	=	211485

mean	code	Model 1 PG1b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.133	0.102	0.126
	2	Other business manager, arts/media/sport	0.182	0.159	0.177
	3	Tradesman/woman, clerks, sales and servi	0.235	0.226	0.233
	4	Machine operators, hospitality staff, as	0.185	0.204	0.189
	8	Not in paid work	0.265	0.309	0.275
sd	1	Senior management	0.00	0.00114	0.00000
	2	Other business manager, arts/media/sport	0.00	0.00149	0.00042
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00280	0.00067
	4	Machine operators, hospitality staff, as	0.00	0.00245	0.00070
	8	Not in paid work	0.00	0.00297	0.00082

Model 2: StudMean + 10 Census variables

Number	of	observed	=	164720
Number	of	imputed	=	44973
Number	of	completed	=	209693

mean	code	Model 2 PG1b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.133	0.098	0.125
	2	Other business manager, arts/media/sport	0.182	0.151	0.175
	3	Tradesman/woman, clerks, sales and servi	0.235	0.220	0.232
	4	Machine operators, hospitality staff, as	0.185	0.210	0.190
	8	Not in paid work	0.265	0.321	0.277
sd	1	Senior management	0.00	0.00155	0.00052
	2	Other business manager, arts/media/sport	0.00	0.00175	0.00048
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00275	0.00063
	4	Machine operators, hospitality staff, as	0.00	0.00208	0.00057
	8	Not in paid work	0.00	0.00239	0.00070

Model 3: StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	164720
Number	of	imputed	=	44902
Number	of	completed	=	209622

mean	code	Model 3 PG1b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.133	0.099	0.126
	2	Other business manager, arts/media/sport	0.182	0.152	0.175
	3	Tradesman/woman, clerks, sales and servi	0.235	0.219	0.232
	4	Machine operators, hospitality staff, as	0.185	0.210	0.190
	8	Not in paid work	0.265	0.321	0.277
sd	1	Senior management	0.00	0.00134	0.00053
	2	Other business manager, arts/media/sport	0.00	0.00106	0.00052
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00247	0.00052
	4	Machine operators, hospitality staff, as	0.00	0.00223	0.00067
	8	Not in paid work	0.00	0.00301	0.00079

Appendix B 4: Results for Level of School_Education (Parent 2)

Model 1: StudMean

Number	of	observed	=	185858
Number	of	imputed	=	25627
Number	of	completed	=	211485

Model 1

mean	code	PG2b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.089	0.099	0.090
	2	Year 10 or equivalent	0.333	0.332	0.333
	3	Year 11 or equivalent	0.077	0.071	0.076
	4	Year 12 or equivalent	0.501	0.499	0.501
sd	1	Year 9 or equivalent or below	0.00	0.00295	0.00047
	2	Year 10 or equivalent	0.00	0.00360	0.00053
	3	Year 11 or equivalent	0.00	0.00195	0.00052
	4	Year 12 or equivalent	0.00	0.00251	0.00032

Model 2: StudMean + 10 Census variables

Number	of	observed	=	185858
Number	of	imputed	=	24863
Number	of	completed	=	210721

Model 2

mean	code	PG2b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.089	0.103	0.090
	2	Year 10 or equivalent	0.333	0.332	0.333
	3	Year 11 or equivalent	0.077	0.073	0.077
	4	Year 12 or equivalent	0.501	0.491	0.500
sd	1	Year 9 or equivalent or below	0.00	0.00157	0.00052
	2	Year 10 or equivalent	0.00	0.00250	0.00042
	3	Year 11 or equivalent	0.00	0.00084	0.00000
	4	Year 12 or equivalent	0.00	0.00237	0.00000

Model 3: StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	185858
Number	of	imputed	=	24816
Number	of	completed	=	210674

Model 3

mean	code	PG2b_School_Educ	Observed	Imputed	Completed
	1	Year 9 or equivalent or below	0.089	0.103	0.090

	2	Year 10 or equivalent	0.333	0.337	0.333
	3	Year 11 or equivalent	0.077	0.072	0.077
	4	Year 12 or equivalent	0.501	0.488	0.500
sd	1	Year 9 or equivalent or below	0.00	0.00226	0.00048
	2	Year 10 or equivalent	0.00	0.00222	0.00042
	3	Year 11 or equivalent	0.00	0.00151	0.00048
	4	Year 12 or equivalent	0.00	0.00244	0.00042

Appendix B 5: Results for Level of NonSchool_Education (Parent 2)

Model 1: StudMean

Number	of	observed	=	161801
Number	of	imputed	=	49684
Number	of	completed	=	211485

mean	code	Model1 PG2b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.196	0.283	0.217
	5	Certificate I to IV (including trade cer	0.409	0.438	0.416
	6	Advanced diploma/Diploma	0.135	0.120	0.132
	7	Bachelor degree or above	0.260	0.159	0.236
sd	4	No non-school education	0.00	0.00247	0.00053
	5	Certificate I to IV (including trade cer	0.00	0.00298	0.00085
	6	Advanced diploma/Diploma	0.00	0.00199	0.00048
	7	Bachelor degree or above	0.00	0.00166	0.00057

Model 2: StudMean + 10 Census variables

Number	of	observed	=	161801
Number	of	imputed	=	48035
Number	of	completed	=	209836

mean	code	Model 2 PG2b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.196	0.291	0.218
	5	Certificate I to IV (including trade cer	0.409	0.441	0.416
	6	Advanced diploma/Diploma	0.135	0.118	0.131
	7	Bachelor degree or above	0.260	0.150	0.235
sd	4	No non-school education	0.00	0.00277	0.00052
	5	Certificate I to IV (including trade cer	0.00	0.00270	0.00063
	6	Advanced diploma/Diploma	0.00	0.00140	0.00048
	7	Bachelor degree or above	0.00	0.00165	0.00048

Model 3: StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	161801
Number	of	imputed	=	47957
Number	of	completed	=	209758

mean	code	Model 3 PG2b_NonSchool_Educ	Observed	Imputed	Completed
	4	No non-school education	0.196	0.291	0.218
	5	Certificate I to IV (including trade cer	0.409	0.442	0.416
	6	Advanced diploma/Diploma	0.135	0.118	0.131
	7	Bachelor degree or above	0.260	0.149	0.235
sd	4	No non-school education	0.00	0.00288	0.00082

5	Certificate I to IV (including trade cer	0.00	0.00162	0.00048
6	Advanced diploma/Diploma	0.00	0.00175	0.00067
7	Bachelor degree or above	0.00	0.00181	0.00048

Appendix B 6: Results for Occupation Group (Parent 2)

Model 1: StudMean

Number	of	observed	=	170888
Number	of	imputed	=	40597
Number	of	completed	=	211485

mean	code	Model 1 PG2b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.193	0.150	0.185
	2	Other business manager, arts/media/sport	0.238	0.210	0.233
	3	Tradesman/woman, clerks, sales and servi	0.254	0.250	0.253
	4	Machine operators, hospitality staff, as	0.248	0.296	0.258
	8	Not in paid work	0.066	0.095	0.072
sd	1	Senior management	0.00	0.00151	0.00032
	2	Other business manager, arts/media/sport	0.00	0.00184	0.00053
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00255	0.00063
	4	Machine operators, hospitality staff, as	0.00	0.00287	0.00053
	8	Not in paid work	0.00	0.00178	0.00032

Model 2: StudMean + 10 Census variables

Number	of	observed	=	170888
Number	of	imputed	=	39036
Number	of	completed	=	209924

mean	code	Model 2 PG2b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.193	0.138	0.183
	2	Other business manager, arts/media/sport	0.238	0.198	0.231
	3	Tradesman/woman, clerks, sales and servi	0.254	0.244	0.253
	4	Machine operators, hospitality staff, as	0.248	0.317	0.261
	8	Not in paid work	0.066	0.104	0.073
sd	1	Senior management	0.00	0.00133	0.00042
	2	Other business manager, arts/media/sport	0.00	0.00211	0.00053
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00263	0.00053
	4	Machine operators, hospitality staff, as	0.00	0.00227	0.00047
	8	Not in paid work	0.00	0.00165	0.00042

Model 3: StudMean + 10 Census variables+ ATSI +Remoteness

Number	of	observed	=	170888
Number	of	imputed	=	38961
Number	of	completed	=	209849

mean	code	Model 3 PG2b_Occ_Group	Observed	Imputed	Completed
	1	Senior management	0.193	0.137	0.183
	2	Other business manager, arts/media/sport	0.238	0.198	0.231
	3	Tradesman/woman, clerks, sales and servi	0.254	0.244	0.252
	4	Machine operators, hospitality staff, as	0.248	0.317	0.261
	8	Not in paid work	0.066	0.104	0.073
sd	1	Senior management	0.00	0.00157	0.00052
	2	Other business manager, arts/media/sport	0.00	0.00237	0.00048
	3	Tradesman/woman, clerks, sales and servi	0.00	0.00247	0.00052
	4	Machine operators, hospitality staff, as	0.00	0.00148	0.00032
	8	Not in paid work	0.00	0.00158	0.00042

Appendix C: Plot of Residuals from OLS School-Level Regression Analysis, 2012 Data based on complete cases, selective school excluded in regression.