TECHNICAL REPORT / AUGUST 2014

Family Occupation and Education Index (FOEI) 2013

Centre for Education Statistics and Evaluation

CENTRE FOR EDUCATION STATISTICS AND EVALUATION WWW.CESE.NSW.GOV.AU

Publication and Contact Details

Centre for Education Statistics and Evaluation 2014,
Family Occupation and Education Index (FOEI) 2013,
Report prepared by L Lu & K Rickard. This report is available at:
http://www.cese.nsw.gov.au/publications/reports-and-presentations

Authors:

Karen Rickard

Statistician | Statistics Unit
Centre for Education Statistics and Evaluation
Strategic Information and Reporting
Office of Education
NSW Department of Education and Communities
T 02 9561 1234
F 02 9561 8055

Dr. Lucy Lu

Leader | Statistics Unit
Centre for Education Statistics and Evaluation
Strategic Information and Reporting
Office of Education
NSW Department of Education and Communities
T 02 9561 8691
F 02 9561 8055

NSW GOVERNMENT | Education & Communities

Contents

List of tables

List of figures

Glossary

ABS	Australian Bureau of Statistics
ACARA	Australian Curriculum, Assessment and Reporting Authority
CESE	Centre for Education Statistics and Evaluation
DEC	Department of Education and Communities
ERN	Enrolment Registration Number
FOEI	Family Occupation and Education Index
ICSEA	Index of Community Socio-Educational Advantage
NAPLAN	National Assessment Program – Literacy and Numeracy
NIASRA	National Institute for Applied Statistics Research Australia
PSFP	Priority Schools Funding Program
RAM	Resource Allocation Model

Overview

The Family Occupation and Education Index (FOEI) is a school-level index of educational disadvantage related to socio-economic background. It has been selected as the basis of the equity loading for socio-economic background in the Department's new Resource Allocation Model.

FOEI is constructed from parental education and occupation information collected from student enrolment forms and recorded in the Enrolment Registration Number system. FOEI uses a statistical regression model to produce a weighted combination of school-level parental education and occupation variables based on the extent to which each variable uniquely predicts average school performance.

This technical report details the construction of FOEI in 2013, and discusses the statistical techniques used to deal with two particular methodological issues – missing parental data and regression outliers – as recommended by statistical experts at the National Institute for Applied Statistics Research Australia through a commissioned review of the FOEI methodology.

Firstly, a multiple imputation methodology was employed to reduce bias due to missing parental data. Multiple imputation produces plausible values for the missing data based on the relationships between existing parental data and a series of other related variables. External validity evidence and a simulation study indicate that multiple imputation has reduced the bias due to missing data to an extent.

Secondly, robust regression analysis was used to reduce the influence of outliers on the regression model. Robust regression is an efficient technique for dealing with outliers by reducing their weight in the regression analysis, leading to more accurate and reliable regression parameters.

Since the introduction of FOEI as the basis of funding to schools, there has been a sizeable reduction in the level of missing data. Further work with schools is planned to continue to reduce the amount of missing data and improve the quality of parental data.

An ongoing program of work is also planned to review and further validate the FOEI methodology to ensure FOEI is the fairest and most robust measure possible of the relative socio-economic disadvantage of NSW government schools.

1 Introduction

Low socio-economic background is a recognised source of educational disadvantage and is associated with lower levels of student achievement. Targeted funding for schools serving low socio-economic communities has existed for many years to address this aspect of educational disadvantage. In 2012, the NSW Department of Education and Communities (DEC) proposed a new Resource Allocation Model (RAM) which included an equity loading for socio-economic background. For this loading, a fair and robust measure of the relative socio-economic disadvantage of each school community was needed.

Until 2013, NSW government school funding for socio-economic disadvantage was provided through the Priority Schools Funding Program (PSFP) and based on a sample survey of parents every four years from schools that chose to participate (about 50 per cent of NSW government schools). This survey methodology is no longer appropriate as much of the information previously collected via the survey is now available from student enrolment forms and recorded in the Enrolment Registration Number (ERN) system. Unlike the PSFP survey, parental information in ERN is potentially available for all students, for all schools, and reflects the current student population.

Another possible candidate is the annual Index of Community Socio-Educational Advantage (ICSEA) developed by the Australian Curriculum, Assessment and Reporting Authority (ACARA) for most Australian schools. ICSEA is constructed as the basis for comparing school performance on the National Assessment Program – Literacy and Numeracy (NAPLAN), and combines several aspects of educational disadvantage (i.e., socio-economic, Aboriginality, remoteness). However, as each of these aspects of disadvantage is funded separately in the RAM, ICSEA is not appropriate as a basis for the equity loading for socio-economic background. In addition, other issues related to the timing and scope of ICSEA preclude its use as the basis for equity funding for NSW government schools.

The Family Occupation and Education Index (FOEI) is a school-level index of educational disadvantage related to socio-economic background. It is derived from information collected from enrolment forms and is constructed to reflect the relative disadvantage of each school's student population relative to other government schools in NSW.

FOEI was originally developed by DEC in 2009 to inform the internal review of ICSEA scores for NSW government schools. In 2012, FOEI was proposed as the basis of the equity loading for socio-economic background in the RAM. To ensure FOEI was fit for purpose as the basis for funding, DEC commissioned statistical experts at the National Institute for Applied Statistics Research Australia (NIASRA) at the University of Wollongong to review the FOEI methodology and to provide technical advice on two particular methodological issues (missing parental data, and regression outliers) which can impact on the index validity. As a result of the review[1], several enhancements were made to the FOEI methodology in 2013.

This technical report details the construction of FOEI in 2013. In the following section, the conceptual issues related to the construction of a measure of socio-economic background are discussed. Section 3 of the report provides further details on the parental education and occupation data used for FOEI, and discusses the problem of missing data. Section 4 then explains the statistical technique used to deal with the missing data and presents the results from the application of this technique. The construction of FOEI scores for schools is presented in Section 5 along with the issue of regression outliers and the statistical technique selected to deal with them. Section 6 details the construction of the FOEI student quarter distribution which is used in conjunction with FOEI scores for funding calculations. Finally, Section 7 outlines the ongoing and further work planned for FOEI.

[1] The NIASRA report is available on the CESE website at: http://cese.nsw.gov.au/reports/research-reports/item/41-methodological-advice-on-family-occupation-and-education-index

2 Conceptual issues

Socio-economic background is a multi-dimensional concept that can potentially be measured by a broad range of social and economic indicators. The key indicators of socio-economic background for children and young people typically reported in the literature include parental education levels, parental occupation status and household income/wealth (e.g. Ainley, Graetz, Long & Batten, 1995; Marks, McMillan, Jones & Ainley, 2000; Butler 2012).

This section discusses some of the issues commonly associated with constructing measures of socio-economic background.

2.1 Direct measures vs area-based measures

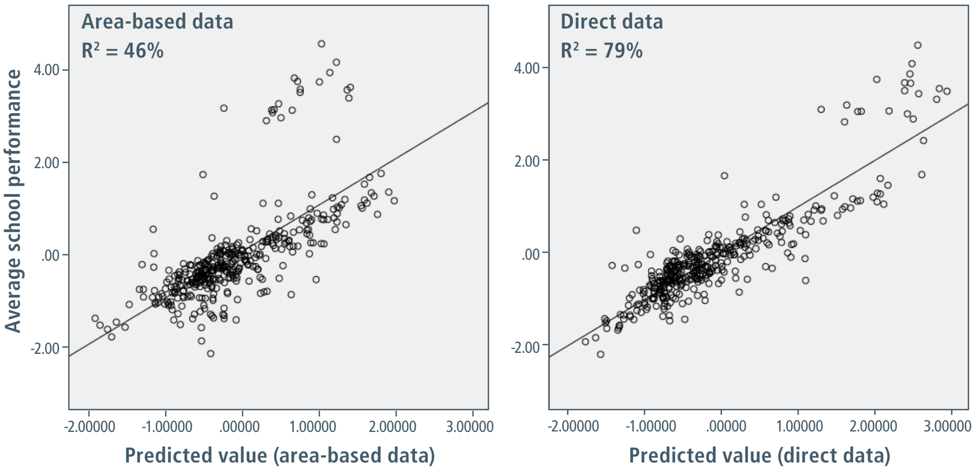
Measures of socio-economic background can be based on either direct data for individual students (direct measures), or indirectly, on the average characteristics of persons residing in the area in which students live (area-based measures). Educational research comparing direct measures and area-based measures demonstrates that area-based measures are not as strongly related to educational outcomes as direct measures and that area-based measures give rise to significant misclassification errors (Ainley et al., 1995; Marks et al., 2000; Lim & Gemici, 2011). The equating of students' characteristics to the average for the area they live in (referred to in the literature as the "ecological fallacy") ignores the significant heterogeneity that exists in many areas. Therefore, indirect area-based measures of socio-economic background are likely to be biased for many students and schools, relative to the extent to which students' families are atypical of the population of the area in which they live.

Area-based measures tend to be used when direct information on individual students is not available. When direct data is available, researchers generally argue that it is superior to area-based measures, especially for targeting resources for students from low socio-economic backgrounds (Lim, Gemici, Rice & Karmel, 2011). NSW DEC has collected nationally consistent information on parental education levels and occupation status via student enrolment forms for a number of years, and direct information is now readily available for the majority of students in NSW government schools. Consistent with research findings, previous DEC analysis comparing direct and indirect data has demonstrated that direct parental education and occupation data are a much stronger predictor of average school performance than indirect data which includes average household income data, for NSW government schools[2], especially for secondary schools, as depicted in Figure 1. Therefore, direct data is the preferred basis of a measure of socio-economic background for funding purposes.

Figure 1:

Scatterplots of average school performance with predicted values using area-based data and direct data, for all secondary schools

Source: 2012 student enrolment data and 2011 average school NAPLAN performance



[2] For this analysis, using 2012 student enrolment data and 2011 average school NAPLAN performance, direct parent data accounted for 79 per cent of the variance in average school performance across all secondary schools, compared to only 46 per cent for indirect, area-based data. For primary schools, direct parent data accounted for 66 per cent of the variance in average school performance compared to 57 per cent for indirect, area-based data.

One issue associated with the use of direct information, however, is the problem of missing parental data. The extent and pattern of missing parental data is discussed further in Section 3, and the technique selected to deal with missing data, as recommended by the NIASRA review, is detailed in Section 4 of this report.

2.2 Composite vs single factor measures

A measure of socio-economic background can be based on a single indicator, such as occupation status, or by combining several indicators to form a composite measure or index. Marks et al. (2000) note that composite measures tend to correlate more strongly with school performance than single indicators. Further, combining different aspects of socio-economic background generally increases the reliability of the resulting measure relative to single indicator measures.

In this regard, the socio-economic background indicators potentially available for constructing a composite measure for NSW government schools include parental education and occupation. Although household income/wealth is an important aspect of a socio-economic construct, this information is not available as it is not collected from student enrolment forms due to the sensitive nature of this type of information. However, research suggests that educational disadvantage in Australia is more strongly related to social factors that are reflected by parental education and occupation, than to material or economic factors such as income or wealth (Marks et al., 2000).

Given these considerations, a composite measure, based on the parental education and occupation information collected from student enrolment forms, has been selected as the basis of a socio-economic background measure for NSW government schools.

2.3 Statistical methodology to construct composite measures

The construction of a valid and reliable composite measure requires a sound methodology to weight and combine the individual indicators.

Several methodologies are possible. Indices such as the Socio-Economic Indices for Areas (SEIFA) developed by the Australian Bureau of Statistics are constructed using principal components analysis (PCA). PCA transforms a large number of correlated variables into one or more composite variables that explain the maximum amount of variance across the original variables. The weighting of each variable in the composite measure reflects the relative size of the contribution each variable makes to the composite measure. As such, the variable weightings have no substantive meaning beyond the statistical PCA model.

An alternative method is to weight the single indicators by the extent to which they are associated with an outcome variable of interest. As equity funding targets educational disadvantage associated with low socio-economic background, it makes sense to combine socio-economic indicators in a way that best predicts that particular educational disadvantage. Regression analysis provides the methodology for empirically combining socio-economic indicators by determining the relative importance (weight) of each indicator in predicting educational outcomes. A drawback to this approach, however, is that empirically derived weights can change over time if the relationship between socio-economic background and educational outcomes changes. This exposes FOEI to 'construct irrelevant variance' as it is potentially influenced by factors not related to the construct itself.

A third option is to use a measurement approach, such as the Rasch framework, to develop a student-level scale for socio-economic status (SES). This approach conceptualises SES as an underlying latent construct that is reflected in the levels of parent education and occupation. Such an approach is also able to deal with missing data. However, this approach would require a major revision to the methodology for generating the school-level measures, and there was insufficient time to explore this approach for FOEI 2013. As part of the ongoing review of the FOEI methodology, this approach will be further investigated.

FOEI for 2013, therefore, was constructed using a regression methodology to weight and combine school-level parental education and occupation indicators based on the extent to which each parental indicator predicts average school performance, as measured by NAPLAN results. This methodology is similar to that underpinning both ICSEA (until 2012) and the old PSFP index. Details of the regression methodology used for FOEI are provided in Section 5.

2.4 Summary

Based on the above considerations, FOEI has been constructed as a composite measure based on readily available direct indicators of socio-economic background (i.e., parental education and occupation) that are combined using a regression analysis that weights each indicator by the extent to which it predicts average school NAPLAN performance.

3 Parental background data

FOEI uses parental education and occupation information provided by parents/carers on student enrolment forms and recorded in the Department's Enrolment Registration Number (ERN) system. This information includes parents'/carers':

- level of school education (the highest year of secondary education a parent/carer has completed)
- educational qualifications (the highest qualification attained by a parent/carer in any area of study other than school education)
- occupation group (the grouping of occupations, based on skill level, that includes the main work undertaken by the parent/carer)

The response categories for these questions on student enrolment forms, as shown in Table 1, are governed by nationally agreed data specifications documented in the Australian Education, Early Childhood Development and Youth Senior Officials Committee (AEEYSOC) Data Standards Manual.

Table 1:

Parental education and occupation response categories

Question	Response categories
School education	- Year 9 or equivalent or below - Year 10 or equivalent - Year 11 or equivalent - Year 12 or equivalent
Educational qualifications	- No non-school qualification - Certificate I to IV (including trade certificate) - Advanced diploma/Diploma - Bachelor degree or above
Occupation group	- Not in paid work in last 12 months - Machine operators, hospitality staff, assistants, labourers and related workers - Tradespeople, clerks and skilled office, sales and service staff - Other business managers, arts/media/sportspersons and associate professionals - Senior management in large business organisation, government administration and defence, and qualified professionals

For use in FOEI, the parent response categories are coded from 1 to 4 (or 5 for occupation group) where 1 represents the lowest category, and 4 (or 5) represents the highest category. The response categories for the two education questions are clearly ordered categories, and therefore represent ordinal variables. The response categories for the occupation question may not completely represent ordered categories, as the 'not in paid work' category could be substantively different in meaning to the other four occupation groups which reflect differing levels of occupational skill and status. The treatment of the occupation variable is discussed further in the relevant sections of this report.

3.1 Data extraction

Parental background[3] data is extracted from ERN in April each year for both parents/carers (if available) for every student[4] enrolled at each school. This means that parental information for siblings enrolled at the same school is extracted for each child and contributes to the school parental background dataset as many times as there are children from the same family enrolled at the school.

Further, some students can attend more than one school. This situation usually arises when a student at a regular primary or secondary school (the home school) is temporarily placed in a special learning or behaviour program at another school for a period of time. In general, students are only counted for FOEI at their home schools. However, for some schools, such as certain Schools for Specific Purposes (SSP), the majority of students attending remain officially enrolled at their home school, rather than the SSP they are temporarily attending. Therefore, in order to produce a reliable and stable FOEI for those SSPs (which generally cater to a small number of students), the parental background data for the students currently attending, but who remain enrolled at their home school, is also included in the data extraction for the SSP.

For 2013, the extraction date of student data from ERN was April 12th. The total number of students extracted was 765,365, of which 649,159 students (85 per cent) had information for two parents and 116,206 students (15 per cent) had information for only one parent. A total of 5,132 students were counted at SSPs, with 4,250 students officially enrolled in an SSP and a further 882 students (17 per cent of all students counted at SSPs) attending an SSP but enrolled at other schools being counted against both their home school and the SSP.

As shown in Table 2 the majority of parents designated as Parent 1 are mothers and the majority of parents designated as Parent 2 are fathers. Overall, 3 per cent of parent records extracted were carers with relationships other than parents or step-parents to students.

Table 2:

Number and percentage of parents by relationship to students

Source: 2013 enrolment data

Note: Percentages may not sum to 100% due to rounding

Relationship to student	Parent 1		Parent 2		Total Parent Records	
	Count	Percent	Count	Percent	Count	Percent
Mother	718,169	93.8%	9,992	1.5%	728,161	51.5%
Father	26,229	3.4%	587,162	90.4%	613,391	43.4%
Step-parent	303	0.0%	31,309	4.8%	31,612	2.2%
Other	20,664	2.7%	20,696	3.2%	41,360	2.9%
TOTAL	765,365	100%	649,159	100%	1,414,524	100%

3.2 Missing data

Due to the optional nature of parental education and occupation questions on enrolment forms, parental background data is not complete for all students. The rate of missing data varies for each of the parental background variables. Figure 2 shows that parental school education has the lowest level of missing data (9 per cent of parents in 2013) and that educational qualifications has the highest level of missing data (20 per cent), followed closely by occupation group (18 per cent). Figure 2 also shows that rates of missing data have been steadily improving over the last four years. From 2010 to 2013 missing data rates have decreased by around 10 percentage points for each parental background variable, in line with strategies encouraging schools to improve the completeness of student and parent data in ERN[5].

[3] For convenience, the term 'parental background' is used to refer jointly to parental education and occupation.

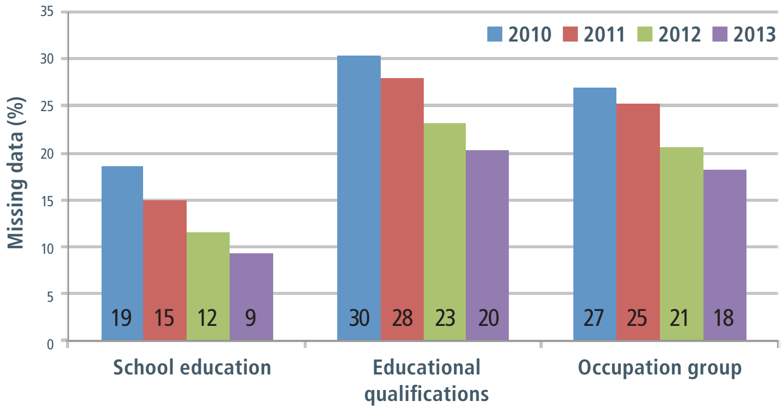
[4] This includes preschool student enrolments for infants/primary schools with preschools attached.

[5] Preliminary analysis of 2014 missing data indicates a further steep drop in missing data rates since the introduction of FOEI for funding, with missing school education down to 6 per cent of parents, missing educational qualifications down to 13 per cent, and missing occupation group down to 11 per cent of parents.

Figure 2:

Percentage of parent records with missing data for each parental background variable

Source: 2010-2013 enrolment data



At the school level, however, there is considerable variation in rates of missing data, with some schools experiencing missing parental information rates well over 50 per cent.

At the student level, most students have at least some parental background information. Students with two parents have up to 6 parental background variables (3 for each parent) whereas students with one parent have a maximum of 3 parental background variables. As shown in Table 3, only 3.6 per cent of students had missing data across all parental background variables (i.e., either all 6 variables for students with two parents, or all 3 variables for students with one parent).

Table 3:

Percentage of students by level of missing parental data

Source: 2013 enrolment data

Note: Students with 1 parent where all 3 parent background variables are missing are reported against "All variables missing"; "na" indicates not applicable.

Level of missing parental data	Percentage of students with:		
	2 parents	1 parent	Overall
All parental data complete	61.2%	58.2%	60.8%
1 variable missing	13.9%	24.8%	15.6%
2 variables missing	10.6%	9.2%	10.4%
3 variables missing	6.5%	na	5.5%
4 variables missing	3.8%	na	3.3%
5 variables missing	1.1%	na	0.9%
All variables missing	2.8%	7.8%	3.6%

Strategies to support schools to improve the completeness and quality of parental information recorded in ERN have been and will continue to be implemented, but due to the optional nature of these questions on student enrolment forms, parental information is not expected to be complete for all parents.

3.2.1 The problem with missing data

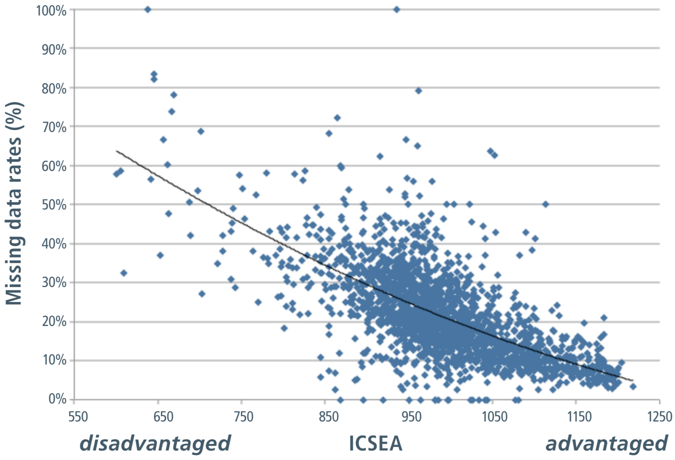
Missing parental information could cause bias in the construction of FOEI if it is not randomly missing. Analysis of missing data patterns for NSW government schools indicates that data is not missing completely at random (MCAR): schools with higher levels of educational disadvantage have higher rates of missing data[6] than schools that are educationally advantaged, as shown in Figure 3.

[6] At the school level, rates of missing data are calculated by dividing the number of parents with missing values for a particular background variable, by the total number of parents of students enrolled at the school.

Figure 3:

Missing data rates for parental educational qualifications by school ICSEA score

Source: 2013 enrolment data; 2012 ICSEA values



The non-random nature of missing data is also demonstrated by a comparison of the rates of missing data in one parental variable for parents who did provide a response for another parental variable. For parents who provided information on their level of school education, the rates of missing data for educational qualifications and occupation group are shown in Table 4. For example, educational qualifications are missing for 24 per cent of parents who indicated that their level of school education was Year 9 compared to only 10 per cent of parents who indicated that their level of school education was Year 12. A similar pattern is evident for missing occupation group. If data was missing completely at random, we would expect the percentages to be similar across the levels of school education.

Table 4:

Rates of missing data in educational qualifications and occupation groups for parents who did provide information on their level of school education

Source: 2013 enrolment data

Level of school education	Percentage of parents with missing data for:	
	Educational qualifications	Occupation group
Year 9	24%	24%
Year 10	15%	20%
Year 11	14%	18%
Year 12	10%	9%

The non-random nature of missing data affects both the estimation of regression coefficients in the regression model and the FOEI scores for individual schools. It means that FOEI scores could be biased for some schools if students with missing data are excluded from the calculation. Excluding students with missing data effectively assumes that those students' backgrounds are equivalent to the average for the school. However, as the analysis of missing data patterns suggests, this is unlikely to be the case. Therefore, it is important to find a way to handle the missing data to reduce its potential bias on school FOEI scores.

4 Using multiple imputation to reduce bias due to missing data

Multiple imputation (MI) is a practical and principled statistical method for handling missing values (Little & Rubin, 1987; Schafer, 1999; Dong & Peng, 2013). It produces multiple sets of plausible values for the missing data, which enables valid estimates of uncertainty (or error) that account for the variability within each 'completed' dataset as well as the variability across the multiple datasets due to the imputation.

Of the several different multiple imputation techniques available, 'multiple imputation by chained equations' (MICE) is a model-based technique recommended by NIASRA during the review of FOEI (NIASRA, 2013). This approach uses the relationships between parental background variables and other auxiliary variables for students where data is available to generate multiple plausible values[7] for students where parental information is missing.

[7] The plausible values represent random draws from the posterior predictive distribution (i.e., the distribution of predicted values conditional on the observed data) (White, Royston & Wood, 2011).

MICE is a widely adopted and flexible approach that allows the full use of the observed data from many variables (White, Royston & Wood, 2011). Features of MICE include:

- MICE is a practical approach to creating imputed values for missing data items based on a set of imputation models, one model for each variable with missing values.
- It uses an iterative process for convergence to produce each set of imputed values.
- Since missing data in each variable is imputed using its own imputation model, it does not assume a multivariate normal distribution and it can handle different variable types.

One key assumption of all multiple imputation methodologies is that conditional on the observed data the non-observed data are "missing at random" (MAR). This assumption means that the probability of a response being missing is not related to the actual value of the missing response, after controlling for observed variables. The MAR assumption is not as restrictive as that of "missing completely at random" (MCAR), but it is often difficult to verify that the MAR assumption holds and that the missing data is not in fact "missing not at random" (MNAR). Including a large number of explanatory variables in the imputation model is one way to help satisfy the MAR assumption (White et al., 2011).

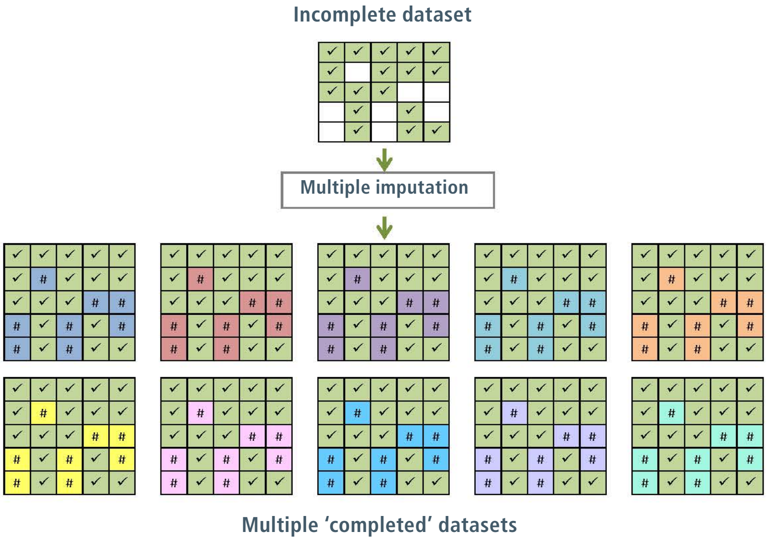
4.1 Using MICE to impute missing parental data for FOEI

The features of MICE described above are important for the FOEI imputation process which requires different imputation models for different types of parental variables. As previously discussed, the two education variables are ordinal variables as the response categories for these questions can be considered ordered categories. Imputation of missing data for these variables therefore requires an ordered logistic regression model. On the other hand, the categories of the occupation variable may not completely represent ordered categories, as the 'not in paid work' category could be substantively different in meaning to the other four occupation groups which reflect differing levels of occupational skill and status. Therefore, the occupation variable is best treated as a categorical variable requiring a multinomial logistic regression model for the imputation of missing data.

Multiple imputation using MICE was carried out using Stata 12 software. A total of ten imputations[8] were performed, resulting in ten 'completed' datasets, as depicted in Figure 4, each consisting of all the non-missing data plus one set of 'plausible values' for the missing data.

Figure 4:

Schematic diagram of the multiple imputation process



[8] The number of imputations was set to 10 consistent with the FOEI methodology review analysis conducted by NIASRA (NIASRA, 2013). The early literature on multiple imputation recommended that 5-10 imputations were generally sufficient to achieve efficiency in estimation (Schafer, 1999). More recently, a number of researchers have recommended a greater number of imputations (at least equal to the proportion of cases with missing data) to achieve more accurate and stable standard errors for estimates and improved statistical power (Allison, 2012; Graham & Olchowski, 2007; White et al., 2011). It is intended in future years to investigate the impact of using a larger number of imputations.

The following sections provide further information on aspects of the multiple imputation process used to impute values for missing parental data in 2013.

4.2 Variables included in the imputation models

The missing data literature (Moons, Donders, Stignen & Harrell, 2006; Kenward & Carpenter, 2007; White et al., 2011) recommends the use of other relevant variables in the imputation process that are:

- correlated with the variables containing missing data, and/or
- predictive of the level of the 'missingness' of the data, and/or
- used in the final analytical model[9], including the dependent variable of that model

On this basis the variables included in the imputation process for each parental background variable include: the other parental background variables; a set of area-based community variables from the ABS census; student achievement scores; student Aboriginal status; and remoteness of school location. The rationale for including each of these variables in the imputation is discussed in the following sections.

4.2.1 Parental background variables

The student-level dataset extracted from ERN potentially contains information for two parents. Parent 1 variables relate to the education and occupation information for the first parent recorded on student enrolment forms. Parent 2 variables relate to the education and occupation information for the second parent recorded on student enrolment forms. The relationships within and between Parent 1 and Parent 2 variables can then be used as part of the imputation process to impute values for those students where some or all of either Parent 1 or Parent 2 information is missing.

For students with complete data, there are generally moderate relationships within and between parents' education levels and occupation groups, as shown in Table 5.

Table 5:

Spearman correlations between education levels and occupation groups for parent 1 (P1) and parent 2 (P2)

Source: 2013 enrolment data.

Note: N's range between 450,000 and 620,000.

	P1 School education	P1 Educational qualifications	P1 Occupation group*	P2 School education	P2 Educational qualifications	P2 Occupation group*
P1 School education	1					
P1 Educational qualifications	.551	1				
P1 Occupation group*	.324	.499	1			
P2 School education	.426	.392	.206	1		
P2 Educational qualifications	.380	.566	.267	.566	1	
P2 Occupation group*	.321	.416	.382	.426	.591	1

* For this analysis, occupation group has been treated as an ordinal variable with 'not in paid work' treated as the lowest occupational category.

Therefore, to aid the imputation of missing data in one parental variable, the other parental variables provide valuable predictive information.

[9] For FOEI, this is the regression model used to weight and combine parental education and occupation variables based on the extent to which each variable predicts average school performance (see section 5 for further details).

4.2.2 ABS community variables

Area-based community variables from the ABS census, linked to student addresses, are an alternative source of socio-economic information that can potentially provide some degree of prediction for both the level of missing data, and for plausible values for missing data.

A set of 11 variables at the statistical area level 1 (SA1)[10] was selected from the 2011 ABS census on the basis that they contributed to the construction of the ABS SEIFA indices and were considered to be relevant to the FOEI construct. They represent the percentage of people or families in an SA1 across a range of education and occupation variables as well as selected income variables and other indicators of disadvantage such as single parent families. The data were then matched to each student record by geocoding student addresses to the SA1 level. Geocoding was successful for 98.1 per cent of students, hence community information was available for the vast majority of students. However, for those students whose address was unable to be geocoded, missing parental data was unable to be imputed.

The 11 ABS community variables selected for the imputation process included the following:

- Percentage of people with annual household equivalised income under $20,799 (INC_UNDER_20799).
- Percentage of people 15 years and over with advanced diploma or diploma qualifications (TE_DIPLOMA).
- Percentage of people 15 years and over with certificate qualifications (TE_CERT).
- Percentage of people 15 years and over with no post-school qualifications (TE_NONE).
- Percentage of people 15 years and over whose highest level of schooling completed is Year 11 or lower (SE_Y11_LOWER).
- Percentage of people in the labour force who are unemployed (LABOUR_UNEMP).
- Percentage of employed people who work in a skill level 1 occupation (high) (OCC_GRP1_HIGH).
- Percentage of employed people who work in a skill level 4 occupation (mid) (OCC_GRP4_MID).
- Percentage of employed people who work in a skill level 5 occupation (low) (OCC_GRP5_LOW).
- Percentage of occupied dwellings with no internet connection (INTERNET_NONE).
- Percentage of families that are one parent families (SINGLE_P_VS_FAMILY_ALL).

The relationship between the community variables and parent background variables for students where parental data is available is shown in Table 6. Most of the community variables correlate with one or more of the parent variables at 0.2 or higher, and in the direction expected, indicating the capacity of these variables to aid the imputation of missing parental data.

Table 6:

Spearman correlations between parental background variables and ABS community variables

Source: 2013 enrolment data.

Note: N's range between 500,000 and 690,000.

	P1 School education	P1 Educational qualifications	P1 Occupation group*	P2 School education	P2 Educational qualifications	P2 Occupation group*
INCOME_UNDER _20799	-0.189	-0.198	-0.199	-0.171	-0.181	-0.253
TE_DIPLOMA	0.252	0.281	0.222	0.253	0.274	0.300
TE_CERT	-0.172	-0.205	-0.019	-0.252	-0.271	-0.145
TE_NONE	-0.307	-0.386	-0.294	-0.319	-0.394	-0.402
SE_Y11_LOWER	-0.317	-0.345	-0.160	-0.378	-0.390	-0.298
LABOUR_UNEMPL	-0.170	-0.186	-0.228	-0.123	-0.148	-0.250
OCC_GROUP1_HIGH	0.314	0.382	0.325	0.305	0.374	0.427
OCC_GROUP4_MID	0.034	0.000	0.087	-0.013	-0.046	0.027
OCC_GROUP5_LOW	-0.146	-0.218	-0.137	-0.184	-0.250	-0.255
SINGLE_P_VS_ FAMILY_ALL	-0.236	-0.272	-0.216	-0.216	-0.261	-0.283
INTERNET_NONE	-0.260	-0.286	-0.225	-0.262	-0.287	-0.320

* For this analysis, occupation group has been treated as an ordinal variable with 'not in paid work' treated as the lowest occupational category.

[10] SA1s are the smallest geographical unit for the release of ABS census data. They generally have a population of 200-800 persons, with an average of 400 persons.

4.2.3 Student achievement scores

There is a strong recommendation in the literature (Kenward & Carpenter, 2007; Moons et al., 2006; White et al., 2011) that the dependent variable in the final analytical model be included as a variable to aid the imputation process. Studies on multiple imputation show that omitting the dependent variable from the imputation model could lead to bias in the estimates from the analysis model (Moons et al., 2006). As the dependent variable in the FOEI analytical model is NAPLAN achievement, this variable has also been included in the imputation model for missing parental data.

The specific NAPLAN measure used in the FOEI analytical model is an average of students' standardised reading and numeracy scores (this is discussed further in section 5 on the FOEI regression model). Therefore the student level NAPLAN measure used in the imputation process is also the average of each student's standardised reading and numeracy scores, if both are available (or if only one standardised score is available, that score is used).

As shown in Table 7, parental education and occupation correlate with composite student NAPLAN scores at around 0.3, indicating the predictive capacity of NAPLAN to aid in the imputation of missing parental data.

Interestingly, while the correlations of school education and educational qualifications with NAPLAN scores are similar for Parent 1 and Parent 2, the correlation of occupation group with NAPLAN is lower for Parent 1 than for Parent 2. This is likely related to the fact that Parent 1 is largely comprised of mothers who are more likely to stay home to care for young children and are therefore not in paid work. As a number of these mothers are likely to have had high status occupations in the past, the association between occupation group and NAPLAN scores for Parent 1 is weakened when these parents are part of the lowest occupation group.

Table 7:

Spearman correlations of parental background variables with students' NAPLAN scores

Source: 2013 enrolment data.

Note: N's range between 280,000 and 400,000.

Parental background	Correlation with composite NAPLAN scores
P1 School education	0.301
P1 Educational qualifications	0.335
P1 Occupation group*	0.248
P2 School education	0.292
P2 Educational qualifications	0.331
P2 Occupation group*	0.326

* For this analysis, occupation group has been treated as an ordinal variable with 'not in paid work' treated as the lowest occupational category.

4.2.3.1 NAPLAN availability

In any given calendar year, only around one-third of students have NAPLAN scores (i.e., students in Years 3, 5, 7 and 9). If students with missing parental information do not have a NAPLAN score, the missing parental data is unable to be imputed using an imputation model that includes NAPLAN scores. To increase the number of students with NAPLAN scores for the purposes of imputing missing parental data, NAPLAN results from previous years (standardised for the relevant scholastic year and calendar year) for matched students have been included in the dataset. Thus for the FOEI 2013 imputation process (which occurred prior to the 2013 NAPLAN tests), most students in Years 4, 6, 8, 10 in 2013 were matched to their 2012 NAPLAN results, students in Years 5, 7, 9, 11 were matched to their 2011 NAPLAN results and students in Year 12 were matched to their 2010 NAPLAN results, as summarised in Table 8.

Table 8:

NAPLAN data calendar year and cohort used for students enrolled in 2013

*As at the April extraction date, the current year's NAPLAN testing had not yet occurred hence no NAPLAN data is available for Year 3 students; the most recent year's NAPLAN data is 2012 for students enrolled in 2013.

Scholastic Year in 2013	NAPLAN Data Used in Imputation
Kindergarten to Year 3*	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

Incorporating 3 years of NAPLAN data means that well over half of all students in 2013 have achievement scores available to assist in the imputation process, as shown in Table 9.

However, no NAPLAN results are possible for students in Years K-3 as at April 2013 (i.e., NAPLAN results are missing "by design"). NAPLAN results are also not available for any student in Years 4-12 in 2013 who was absent, withdrawn or exempt from NAPLAN testing in the relevant year, or was not attending a NSW government school at the time. Table 9 shows the numbers of students with and without NAPLAN results by year of schooling.

Table 9:

Numbers of students with and without NAPLAN results

Source: 2013 enrolment data and matched NAPLAN results 2010-2012

Note: Students in Year 3 with NAPLAN information are likely to have repeated Year 3 in 2013. Students in Years 1 and 2 with NAPLAN information represent coding errors. Students not part of a NAPLAN cohort in Years 4 to 12 represent those not enrolled in a NSW government school as at the relevant NAPLAN test year.

School Year	Total students 2013	Participated in NAPLAN (results available)	Exempt in Reading & Numeracy	Absent or Withdrawn in Reading & Numeracy	Not part of a NAPLAN cohort	Total without NAPLAN results	Percentage without NAPLAN results
K	70,334	0	0	0	70,334	**70,334**	**100%**
1	68,204	1	0	0	68,203	**68,203**	**100%**
2	65,517	2	2	0	65,513	**65,515**	**100%**
3	62,374	148	22	14	62,190	**62,226**	**100%**
4	61,772	56,636	1,202	1,303	2,631	**5,136**	**8%**
5	60,303	53,536	1,040	1,203	4,524	**6,767**	**11%**
6	59,858	55,851	1,140	1,272	1,595	**4,007**	**7%**
7	52,096	45,349	1,002	1,064	4,681	**6,747**	**13%**
8	53,190	48,935	891	1,606	1,758	**4,255**	**8%**
9	53,945	48,524	916	1,385	3,120	**5,421**	**10%**
10	56,158	49,112	941	3,098	3,007	**7,046**	**13%**
11	53,927	45,268	759	1,820	6,080	**8,659**	**16%**
12	42,898	36,197	641	823	5,237	**6,701**	**16%**
Ungraded	4,789	0	1	0	4,788	**4,789**	**100%**
Total	**765,365**	**439,559**	**8,557**	**13,588**	**303,661**	**325,806**	**43%**

4.2.3.2 Options when NAPLAN results are not available

One option for the imputation process for missing parental background data when NAPLAN results are not available is to use an imputation model that also imputes for missing NAPLAN results in the chained equations procedure. However, the same imputation model cannot be used for all students since the missingness for Years K-3 students is by design, hence it is not appropriate to impute the NAPLAN scores for these particular students (NIASRA, 2013). Furthermore, there would be a potential lack of acceptance by stakeholders, especially if NAPLAN results were to be imputed for students in Years K-3 who have never been part of the NAPLAN testing process.

The alternative to imputing for missing NAPLAN results for all students is to separate students in Years K-3 for the purpose of imputing the parental background variables and use a different imputation model – one that does not include NAPLAN results.

The final decision involved a combination of both options. To maximise the use of NAPLAN results to aid the imputation process for the parental background variables, an imputation model that imputes for missing NAPLAN results was used for students who were absent or withdrawn from NAPLAN testing, as these students were at least part of a NAPLAN cohort. However a different imputation model was used that did not include NAPLAN achievement scores for students in Years K-3 as at April 2013, as well as for students who were exempt from NAPLAN testing (due to their unique circumstances[11]) and those not part of a NAPLAN cohort due to not being enrolled in a NSW government school at the time. To ensure the imputed parental values were as robust as possible for these students, the imputation was based on the relationships among parental and other auxiliary variables for all students across Years K-12 (i.e., all students were included in the alternate imputation process but missing parental data was only imputed for Year K-3 students, exempted students and those not part of a NAPLAN cohort).

4.2.4 Student Aboriginal status

Aboriginal status is strongly associated with socio-economic disadvantage. Table 10 shows that parents of Aboriginal students are much less likely to have completed Year 12, hold a diploma or bachelor degree or above, or be occupied in the higher level occupation categories, than parents of non-Aboriginal students. Student Aboriginal status is therefore a useful auxiliary variable to aid the imputation of missing parental background data.

Table 10:

Percentage of parents by parental background category by student Aboriginal status

Source: 2013 enrolment data.

Parental background categories		Aboriginal students	Non-Aboriginal students
Percentage of students' parents by level of school education	Year 9 or below	24%	8%
	Year 10	42%	28%
	Year 11	11%	8%
	Year 12	23%	57%
Percentage of students' parents by educational qualifications	No educational qualification	42%	21%
	Certificate I to IV (including trade certificate)	43%	34%
	Advanced diploma/Diploma	8%	15%
	Bachelor degree or above	7%	30%
Percentage of students' parents by occupation group	Not in paid work in last 12 months	36%	18%
	Machine operators, hospitality staff, assistants, labourers	28%	20%
	Tradespeople, clerks, skilled office/sales/service staff	18%	23%
	Other managers, arts/media/sports, associate professionals	10%	21%
	Senior managers, qualified professionals	7%	17%

Student Aboriginal status is coded as:

- Aboriginal and/or Torres Strait Islander Origin (n = 50,581, 6.6 per cent of students)
- Neither Aboriginal nor Torres Strait Islander Origin (n = 712,399, 93.1 per cent of students)
- Not stated/Unknown (n = 2,385, 0.3 per cent of students)

Missing parental data was unable to be imputed for those students with not stated/unknown Aboriginal status.

[11] Students can be exempted from NAPLAN tests if they have significant or complex disability, or if they are from a non-English-speaking background and arrived in Australia less than one year before the tests.

4.2.5 School location

Socio-economic disadvantage is also associated with location. The variable used for location was the MCEECDYA remoteness classification for schools, coded into 3 groups:

- Metropolitan (1,272 schools [58.1 per cent] enrolling 573,152 students [74.9 per cent])
- Provincial (863 schools [39.4 per cent] enrolling 187,720 students [24.5 per cent])
- Remote (including Very Remote) (54 schools [2.5 per cent] enrolling 4,493 students [0.6 per cent])

As all schools are classified by location, there is no missing data in this variable.

Table 11:

Percentage of parents by parental background category and school location

Source: 2013 enrolment data.

Parental background categories		Metropolitan	Provincial	Remote
Percentage of students' parents by level of school education	Year 9 or below	8%	11%	18%
	Year 10	25%	39%	37%
	Year 11	7%	10%	12%
	Year 12	60%	40%	33%
Percentage of students' parents by educational qualifications	No non-school qualification	21%	26%	42%
	Certificate I to IV (including trade certificate)	31%	46%	38%
	Advanced diploma/Diploma	16%	11%	9%
	Bachelor degree or above	32%	16%	12%
Percentage of students' parents by occupation group	Not in paid work in last 12 months	19%	19%	25%
	Machine operators, hospitality staff, assistants, labourers	19%	27%	31%
	Tradespeople, clerks, skilled office/sales/service staff	23%	24%	16%
	Other managers, arts/media/sports, associate professionals	21%	19%	19%
	Senior managers, qualified professionals	19%	11%	8%

Table 11 demonstrates that parents of students living in provincial areas, and remote locations especially, have lower levels of school education, educational qualifications and occupation status than parents of students in metropolitan areas. Therefore, location is also a useful auxiliary variable for imputing missing parental background data.

4.2.6 Other variables considered but not used

Student gender and scholastic year were also considered in discussions with NIASRA as potential variables to include in the imputation. However, there is no evidential basis for linking students' gender or learning stage to their parents' education or occupation.

4.2.7 *Summary of variables used in the imputation*

The final set of variables used in the FOEI imputation models therefore included:

- 3 parental background variables for each parent
- 10 'community variables' derived from the 2011 ABS census for the small geographic area (SA1) relating to each student's address
- each student's latest composite NAPLAN score, for those students who participated in NAPLAN tests during 2010, 2011 or 2012
- student Aboriginal status
- school remoteness classification based on geographical location

4.3 Other considerations

4.3.1 *Students with data missing for all parental variables*

Parental background data is completely missing for 27,420 students (3.6 per cent) in 2013. However, the inclusion of the additional variables in the imputation model, as described above, allows for the imputation of plausible values for the missing parental data for these students. It is expected however, that the variability in the plausible values imputed for these students will be greater across the ten imputed datasets, than would otherwise be the case if some of the parental variables were present.

4.3.2 *Students with information available for only one parent*

Approximately 116,000 students (15 per cent) have information recorded in ERN for only one parent (i.e., the Parent 2 section of the student enrolment form was completely blank). For these students there is no Parent 2 information to inform the imputation of missing values in Parent 1 information, nor is it appropriate to impute Parent 2 information for these students, as most of these students are expected to be from single parent families.

In order to carry out the imputation process appropriately for this group of students, the dataset was split into two: one subset for students with two parent records; the other subset for students with only one parent record.

The imputation models for these 2 subsets of data are therefore slightly different, with the two parent dataset including all 6 parent variables and one parent dataset including only the 3 parent variables available.

4.4 Summary of the imputation process

Following is a summary of the steps used to impute plausible values for missing parental data:

1. Student records were separated into two subsets (i.e., students with one parent record and students with two parent records), and the following imputation processes were carried out separately with each set of records.
2. Students with NAPLAN results, or who were in the NAPLAN cohorts but were absent or withdrawn, were selected and an imputation model applied to impute for missing values in the parent background variables (3 variables for the one-parent records, and 6 variables for the two-parent records) and any missing student achievement scores, using the 3 (or 6) parent background variables, student achievement scores, Aboriginal status, school remoteness and 10 community variables. These records were saved as a separate file.
3. All students were re-selected[12]. An imputation model was run to impute for missing values in the parent background variables (3 variables for the one-parent records, and 6 variables for the two-parent records), using the 3 (or 6) parent background variables, Aboriginal status, school remoteness and 10 community variables (i.e., excluding student achievement scores from explanatory variables).
4. From the results of Step 3, students who were not in NAPLAN cohorts or were exempted from NAPLAN testing, were selected and appended to the results from Step 2.
5. The one-parent and two-parent student records with imputed values for missing data were then combined back into a single file comprising all students.

The final file comprised ten ‘completed’ datasets. Each ‘completed’ dataset represents all the non-missing parental data plus one set of imputed values for the missing data.

4.5 Results of the imputation process

The imputation process successfully imputed values for missing data for the majority of students with missing parental information. However, for 9,090 students (1.2 per cent of all students) missing parental information was not able to be imputed due to missing Aboriginal status, or due to unsuccessful geocoding that meant that ABS community variables were not available for these students (the majority of these students are in small schools in rural areas).

At the school level, the imputation process has resulted in almost all schools (98 per cent) having ‘completed’ data (either observed or imputed values) for 90 per cent or more of their students.

4.5.1 Comparison of the distribution of imputed values across the 10 imputed datasets

The distribution of imputed values for missing data was very consistent across the 10 imputed datasets. Table 12 reports the distribution of imputed values for the 279,786 parents (20 per cent of parents) with missing educational qualifications, as an example. The percentage of parents imputed at each level of educational qualifications varies by no more than 0.3 percentage points across the 10 imputed datasets. A similar result was observed for the distribution of imputed values for parental school education levels and occupation group.

[12] See section 4.2.3.2 for an explanation as to why an imputation model excluding student achievement scores was run with all students.

Table 12:

Distribution of imputed values for parents with missing educational qualifications, across the 10 imputed datasets

	No educational qualification	Certificate	Advanced diploma/ Diploma	Bachelor degree or above
Dataset 1	32.8%	39.9%	12.8%	14.5%
Dataset 2	32.9%	39.9%	12.8%	14.4%
Dataset 3	32.9%	39.8%	12.9%	14.4%
Dataset 4	32.7%	40.0%	12.8%	14.4%
Dataset 5	32.8%	40.0%	12.6%	14.6%
Dataset 6	32.9%	39.9%	12.7%	14.5%
Dataset 7	32.6%	40.1%	12.9%	14.4%
Dataset 8	32.7%	40.0%	12.9%	14.4%
Dataset 9	32.7%	40.1%	12.9%	14.4%
Dataset 10	32.9%	39.9%	12.8%	14.4%
Minimum	32.6%	39.8%	12.6%	14.4%
Maximum	32.9%	40.1%	12.9%	14.6%
Difference	0.3	0.3	0.3	0.2

4.5.2 *Comparison of the distribution of observed and imputed values*

Tables 13 to 15 compare the percentage of parents at each level of school education, educational qualification, and occupation category, for:

- the observed data available from ERN
- the imputed values (averaged across the ten imputed datasets) for parents where relevant information was missing
- the 'completed' distribution of parents once the observed and imputed data have been combined (averaged across the ten imputed datasets)

Table 13:

Percentage of parents by level of school education

Note: 9 per cent of parents had missing school education information in 2013.

School education	ERN 2013 observed	Imputed values for missing data	Completed ERN + imputed
Year 9 or equivalent or below	8.5%	10.3%	8.7%
Year 10 or equivalent	28.5%	30.7%	28.7%
Year 11 or equivalent	7.8%	7.3%	7.7%
Year 12 or equivalent	55.2%	51.6%	54.9%

Table 14:

Percentage of parents by level of educational qualification

Note: 20 per cent of parents had missing educational qualifications information in 2013.

Educational qualification	ERN 2013 observed	Imputed values for missing data	Completed ERN + imputed
No non-school qualification	21.9%	32.8%	24.1%
Certificate I to IV	34.8%	40.0%	35.8%
Advanced diploma/Diploma	14.7%	12.8%	14.3%
Bachelor degree or above	28.6%	14.4%	25.8%

Table 15:

Percentage of parents by occupation group

Note: 18 per cent of parents had missing occupation group information in 2013.

Occupation group	ERN 2013 observed	Imputed values for missing data	Completed ERN + imputed
Not in paid work in last 12 months	19.1%	24.1%	19.9%
Machine operators, hospitality staff, assistants, labourers	20.5%	25.6%	21.4%
Tradespeople, clerks, skilled office/sales/ service staff	22.9%	22.1%	22.8%
Other managers, arts/media/sports, associate professionals	20.6%	16.6%	19.9%
Senior managers, qualified professionals	17.0%	11.6%	16.0%

For all parental variables, the distribution of imputed values is shifted towards the 'lower' categories, as expected. For school education, 10.3 per cent of parents with missing data were imputed to have Year 9 or below as their highest level of schooling, compared to 8.5 per cent of parents who actually reported this information. Similarly, for educational qualifications, 32.8 per cent of parents with missing data were imputed to have no qualifications, compared to 21.9 per cent of parents who reported this information.

The fact that, after imputation, the percentage of parents in the various categories is shifted in the expected direction indicates that the methods and processes used to impute the missing parental values have corrected bias arising from missing data to an extent.

4.6 Validation of multiple imputation

4.6.1 External validation against ABS data

As an external validation of the multiple imputation process, customised data on parents' highest level of school education was obtained from the 2011 ABS census for people in NSW with school-aged children attending government schools. This was compared to the observed data from ERN (in 2013), to the imputed values for missing data, and to the 'completed' data comprising both observed and imputed values[13].

Table 16 shows that for parental school education, relative to the ABS census data, the observed data in ERN shows slightly fewer parents in the lower 2 categories (Year 9/Year 10) and slightly more in the higher 2 categories (Year 11/ Year 12). As described above, the imputation process has assigned a slightly higher proportion of the missing data to the lower 2 categories relative to the observed data (e.g., 10.3 per cent of missing data imputed as Year 9 compared to 8.5 per cent observed in this category). When the observed and imputed data are combined, the resulting percentages have all moved in the 'right direction' to more closely approximate the ABS data. However, there are still fewer parents in the lower 2 categories and more parents in the higher 2 categories relative to the ABS data, if we assume that levels of school education of those parents of students attending a government school have not changed significantly between 2011 and 2013. This shows that the imputation process has helped to reduce some of the bias in the missing data, but may not have been able to completely eliminate it.

Table 16:

Highest level of school education for parents of students in NSW government schools

School Education	ABS census 2011	ERN 2013 observed	Imputed values for missing data	Completed ERN + imputed
Year 9 or equivalent or below	9.4%	8.5%	10.3%	8.7%
Year 10 or equivalent	30.2%	28.5%	30.7%	28.7%
Year 11 or equivalent	6.8%	7.8%	7.3%	7.7%
Year 12 or equivalent	53.5%	55.2%	51.6%	54.9%

4.6.2 External validation against Priority Schools Funding Program (PSFP) survey data

The 2012 PSFP survey data provides a useful benchmark for validating the imputation process, as the response rate achieved by the survey is much higher (around 96 per cent on average) than that for enrolment forms. For this comparison, a 2012 student dataset from ERN was subjected to the same multiple imputation process as used for 2013 data. For the 1,103 schools that participated in the 2012 PSFP survey, the observed parental education data from ERN, and the values imputed for missing data, were compared to the parental education data from the PSFP survey data[14].

The distribution of parental school education levels was surprisingly similar for both the PSFP survey and the observed data from ERN for the same schools. That is, missing parental school education data in ERN for this subset of schools did not result in the expected bias in observed levels of school education. Consequently, the imputation process also

[13] ABS census information was also obtained on the educational qualifications and occupation group of people in NSW with school-aged children attending government schools, however issues related to the coding of missing qualifications, and the coding system used for occupation in the census data, meant the census data was not able to be compared to parental data in ERN.

[14] Occupation data from the PSFP survey and ERN could not be compared as the occupation categories collected from the two sources varied considerably.

resulted in a very similar overall distribution of 'completed' parental school education levels. However, the impact was greater for the distribution of parents across different levels of educational qualifications. As shown in Table 17, the PSFP data showed that, of the 309,413 parents who responded to the educational qualification question on the survey, 41.1 per cent indicated they did not have a post-school qualification. This is almost 8 percentage points higher than the observed 33.5 per cent of the 348,976 parents who were from the same schools and provided a response on the enrolment form. After imputation, this percentage was increased by 2.1 percentage points to 35.6 per cent, closer to the proportion reported through the PSFP survey. While issues affecting the accuracy of self-reported data are different for the PSFP data than for enrolment data, it is considered that the imputation has corrected bias in the missing parental data in the right direction.

Table 17:

Comparison of 2012 observed and imputed parental data with parental responses from the 2012 PSFP survey

	Parents' responses from PSFP survey		Parents' responses in ERN for schools participating in the survey			
	Total parents responded	% of parents	Observed in ERN	Total parents responded	Imputed	Completed ERN + imputed
No non-school qualification	127,214	41.1%	33.5%	116,957	40.4%	35.6%
Certificate I to IV	116,173	37.5%	41.7%	145,436	39.5%	41.0%
Advanced diploma/ Diploma	31,161	10.1%	11.1%	38,566	10.8%	11.0%
Bachelor degree or above	34,865	11.3%	13.8%	48,017	9.4%	12.5%
Total	309,413	100%	100%	348,976	100%	100%

4.6.3 Validation using a simulation analysis

An additional test of the extent to which multiple imputation can reduce bias due to missing data was provided by a simulation analysis. For this analysis, a sample of 50 per cent of student records with completely observed parental data for two parents was selected from the original dataset. Missing data was then introduced to the sample dataset to replicate the patterns of missing data observed in the original dataset (i.e., parents with low levels of education and occupation are more likely to have missing data). The same multiple imputation process was applied to the missing data in the simulation sample dataset generating 10 imputed sample datasets. Results reported below represent averages across the 10 imputed datasets.

Table 18 shows the effects of the missing data, and of applying multiple imputation, on the distribution of parental educational qualifications, as an example. The introduction of missing data had the expected effect of shifting the distribution of observed parental educational qualifications towards the higher categories, with the percentage of parents with a bachelor degree or above increasing from 31.6 per cent to 34 per cent, and the percentage of parents with no non-school qualification decreasing from 20.1 per cent to 18.5 per cent. After multiple imputation was applied, the percentages of parents in each category has shifted back towards the lower categories, more closely reflecting the original percentages in the sample dataset.

Table 18:

Percentage of parents by level of educational qualification

Educational qualification	Original sample dataset	After introduction of missing data (observed only)	After imputation (completed = observed + imputed)
No non-school qualification	20.1%	18.5%	20.1%
Certificate I to IV	33.6%	32.3%	33.3%
Advanced diploma/Diploma	14.8%	15.2%	15.0%
Bachelor degree or above	31.6%	34.0%	31.6%

To examine the effect of multiple imputation at the school level, three FOEI scores were calculated for each school (see section 5 for details of the FOEI calculation process): a ‘true FOEI’ based on the original complete sample data, an ‘incomplete FOEI’ based on the data after missing values were introduced, and an ‘imputed FOEI’ based on the final data after multiple imputation was applied.

Overall, ‘imputed FOEI’ scores were slightly closer to ‘true FOEI’ scores (r=0.97) than were ‘incomplete FOEI’ scores (r=0.94). This was especially so for small schools with less than 20 students, where the correlation with ‘true FOEI’ was noticeably higher for ‘imputed FOEI’ (r=0.93) than for ‘incomplete FOEI’ (r=0.87). Similarly, for schools with higher rates of missing data (more than 25 per cent of parental items missing), ‘imputed FOEI’ scores were also considerably closer to ‘true FOEI’ scores (r=0.81) than were ‘incomplete FOEI’ scores (r=0.61).

Therefore, multiple imputation has reduced the bias associated with missing data, especially for small schools and schools with higher rates of missing data.

4.7 Conclusion

The use of multiple imputation has served to provide plausible values for the majority of students with missing data, and appears to have reduced the bias due to missing data to an extent.

5 Calculation of FOEI scores

This section describes the statistical model used to calculate FOEI scores and the variables used in the model. The issue of outliers is discussed along with the specific regression technique selected to deal with them. The results of the 2013 FOEI calculation are presented along with an analysis of the stability of FOEI.

5.1 The FOEI model

FOEI is a school-level measure that is constructed using a statistical regression model to produce a weighted combination of school-level parent education and occupation variables (i.e., percentages of parents in each education/occupation category) based on the extent to which each variable uniquely predicts average school performance.

Mathematically, the FOEI regression model can be written as:

$$y_s = \beta_0 + \beta' X_s + e_s$$

where for school s:

- y_s is average school performance
- X_s is the vector of school-level parent background variables
- β' is the vector of coefficients for the parent background variables
- β_0 is the intercept or constant term
- e_s is the error or residual term

The predicted values from this regression model are used as the basis of the final FOEI scores.

The following section describes the school-level variables used in the FOEI regression model.

5.2 Variables in the model

5.2.1 Average school performance

The dependent variable in the regression analysis is the school average of the most recent year's NAPLAN results which for the 2013 FOEI are the 2012 NAPLAN results. NAPLAN performance was chosen as the measure of school performance as this is the only comparable performance measure available for the majority of NSW government schools. Other potential performance measures, such as HSC results, are only available for a smaller subset of schools, and reflect the performance of fewer students in those schools, than NAPLAN results.

A composite performance score was calculated for each school based on students' reading and numeracy results, equally weighted. A composite measure of performance has been used as this is generally a more reliable indicator of performance at a particular time than a single test score. Students' reading and numeracy scores were first standardised using the mean and standard deviation of NAPLAN scores for all students in NSW government schools for each test cohort. Each student's standardised scores for reading and numeracy were averaged (if both were available, otherwise, if only one result was available, that was used), and then aggregated to a school-level average performance score for each school.

For primary schools, the performance measure reflects the combined performance of the Year 3 and Year 5 student cohorts from 2012. For secondary schools, the performance measure reflects the combined performance of the Year 7 and Year 9 student cohorts from 2012. For central/combined schools, the performance measure reflects the combined performance of Year 3, Year 5, Year 7 and Year 9 from 2012.

Not all schools have NAPLAN performance scores. Infants schools, senior high schools and a number of SSPs have no students participating in NAPLAN tests. These schools therefore are unable to contribute to the development of the FOEI model; however, once developed, the FOEI model can be applied to these schools' parental background data to generate FOEI scores.

5.2.2 School-level parental variables

The parental data used in the FOEI regression model is a combination of the observed parental data and the imputed values for missing parental data aggregated to school-level percentages of parents[15] in each education and occupation category (for example, the percentage of parents whose highest level of school education is Year 12, the percentage of parents with no educational qualifications, and the percentage of parents in occupation category 'tradespeople').

For each school, the percentage of parental responses in each category for each variable was calculated from:

- The weighted number of students' parents with an observed or imputed response in the given category, divided by
- The total weighted number of students' parents with an observed or imputed response for the relevant parental background question

Two principles guided the calculation of these weighted school-level counts:

- Each student contributes parental background information. Therefore, 2 (or more) students from the same family contribute the same parental information 2 (or more) times.
- Each student's parental background information contributes equally to that of other students. Therefore, parental information for students from one parent families (identified as those with information in ERN for only one parent) is given a weight of 2 in the school-level parent counts. This ensures that the parental background of students in one parent families contributes equally to the parental information for students from two parent families.

[15] The school-level percentages of parents in each education and occupation category include both parents for students in two parent families. Other options previously explored for combining the information from two parents for each student to generate school-level percentages are provided in the Appendix.

In total, there are 13 school-level parental variables obtained from the total number of categories across the three parental background variables (i.e., 4 + 4 + 5). Each of these variables is a continuous variable as the quantity is a percentage of parental responses in the given category. However, each subset of variables, relating to each of the three original parental background questions, sums to 100 per cent, such that any one of the variables in each subset can be defined by subtracting the sum of the others from 100 per cent. Therefore, it is necessary to omit one of the variables in each subset from the regression modelling.

Which variable is omitted from each subset has no effect on the predicted values from the regression modelling. The impact is on the interpretation of the resulting regression coefficients, as these represent the size and direction of the effect of each variable on school performance, over and above the effect of all other variables, relative to the omitted variable (also referred to as the reference category).

As FOEI is designed to represent relative levels of school socio-economic disadvantage, it was decided to omit the variable representing the highest category in each subset of variables. The resulting coefficients for the 10 included variables therefore represent the degree of educational disadvantage associated with lower levels of parental education and occupation.

5.3 Analysis considerations following multiple imputation

The school-level parental variables were constructed for each of the ten 'completed' datasets (each containing the observed data plus one set of imputed values for missing data).

Analysis of data following multiple imputation typically involves applying the analysis model to each imputed dataset, and then averaging the resulting parameters (i.e., regression coefficients). In the process, an estimate of the error associated with each parameter can be generated, that includes the uncertainty in both the analysis model estimation and the imputation itself.

There are a number of difficulties associated with this approach for FOEI. Firstly, the imputation is carried out at the student level, but the analysis is carried out at the school level, using aggregated student data. Secondly, the primary outcome of interest is the predicted value for each school, not the regression coefficients. Thirdly, the usual analysis process would lead to significant complexity in reporting to school principals as the resulting FOEI score would be based on ten sets of school-level parental data rather than one. In the interests of transparency and stakeholder acceptance, the decision was made to create one set of school-level parental background data, by averaging the school-level percentages of parents in each education and occupation category across the ten imputed datasets. These average percentages can then be used in the regression analysis to determine each school's FOEI score. Using this method, the links between the school-level data, the regression weights (i.e., estimated coefficients) and the school FOEI score can be more easily and clearly explained to school principals.

This change to the usual analysis method following multiple imputation was accepted by NIASRA during the review of the FOEI methodology (NIASRA, 2013). A comparison of the two methods is planned as part of the ongoing review of the FOEI methodology to examine any practical differences in the final regression weights and resulting FOEI scores.

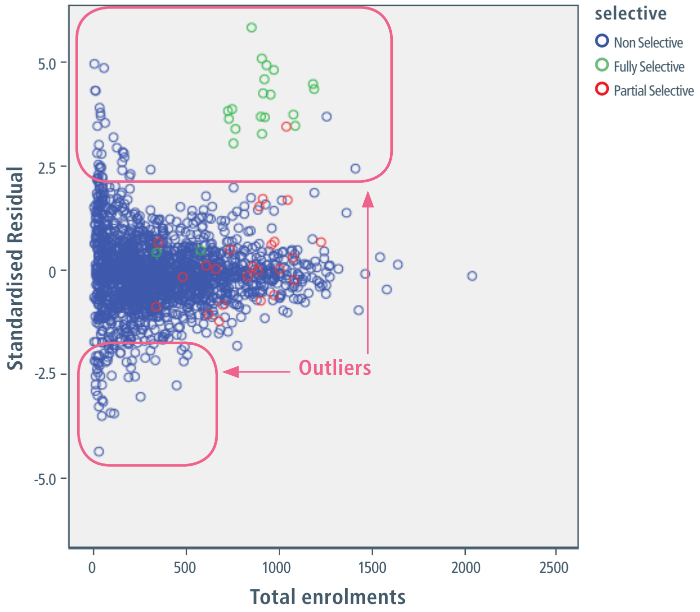
5.4 Regression techniques and the problem of outliers

FOEI was originally developed using an ordinary least squares (OLS) regression technique similar to the development of ICSEA (for 2008-2012) and the PSFP index. However, a number of schools were identified as outliers (i.e., they had standardised residuals greater than 2), indicating that they did not show the same relationship between parental background characteristics and performance observed for the majority of other schools. As indicated in Figure 5, these schools were predominantly academically selective schools and a number of very small schools.

Figure 5:

Scatterplot of standardised residuals from OLS regression against total enrolments

Source: Results from school-level regression analysis based on 2012 enrolment data and average school NAPLAN results



To reduce the influence of outliers and improve the stability of the FOEI regression model from year to year, all selective schools and schools with total enrolments less than 100 were excluded from the regression modelling in previous years. However, as indicated in Figure 5, and elaborated in the NIASRA report (NIASRA, 2013, section 2.3) only some small schools are outliers; the majority of small schools fit the regression model well and should not be excluded from the analysis.

Nevertheless, outliers need to be dealt with as they can significantly affect the regression model coefficients by 'dragging' the regression line towards the outlying values. This in turn affects the predicted values from the regression analysis for all schools. Instead of automatically excluding academically selective schools and all small schools from the regression modelling, a number of alternative regression techniques that can deal with outliers were explored.

5.5 Alternative regression techniques

Alternative regression techniques for dealing with outliers were investigated in conjunction with NIASRA as part of the FOEI review. The alternative techniques considered included weighted least squares (WLS) and robust regression (RR).

Weighted least squares is a technique for pre-determining the weight or influence of individual cases (i.e., schools) in the formulation of the regression model. As the relationship between parental background factors and school performance is weaker, or more variable, for schools with fewer students, weighting by the size of the school NAPLAN cohort was considered one potential option for minimising this effect. That is, schools with a large number of students with NAPLAN results receive greater weight as the relationship between parental background and performance is generally more consistent and reliable, than schools with few NAPLAN student results. As this approach does not deal with the issue of academically selective schools, it is still necessary to manually exclude these schools from the WLS regression analysis.

Robust regression is a technique that is designed to be less affected by violations in the assumptions of OLS regression and is especially useful for reducing the influence of outliers by reducing their weight in the regression analysis. It offers a compromise between excluding outliers entirely from the analysis and including all the data points and treating them equally. The robust regression implemented in Stata (rreg procedure) includes an initial step that removes high-leverage outliers (based on Cook's D) and then uses an iterative M-estimator algorithm (Huber followed by bisquare) to estimate weights for each observation (i.e., school) related to the size of the residuals.

5.5.1 *Summary of a comparison of alternative regression techniques*

A comparison of OLS, WLS and RR was conducted in early 2013 in conjunction with statistical experts at NIASRA, and was based on students with complete parental data in the 2012 student dataset. A summary of the results of this analysis, and the conclusions and advice provided by NIASRA are contained in their report (NIASRA, 2013) available on the CESE website.

In brief, the analysis concluded that robust regression was the most effective method to minimise the effect of outliers on the final regression model. This method produced similar regression coefficients and predicted values to an OLS method where outliers were manually removed, and hence provides a more efficient method for calculating FOEI scores. As expected, schools assigned a zero weight in the robust regression analysis were largely very small schools, and the academically selective schools.

The WLS method (weighting by NAPLAN cohort size) was found to overcompensate for the effect of school size. By imposing a smaller weight for all schools with few NAPLAN students, the influence of the majority of these schools, which did fit the regression model quite well, was unjustifiably reduced. WLS also resulted in regression coefficients and predicted values which were not as well aligned with those from either RR or an OLS method with outliers manually excluded.

Robust regression, as implemented in Stata 12 software, was therefore selected as the regression technique for the construction of the 2013 FOEI scores.

5.6 Results of the 2013 FOEI regression analysis

Table 19 presents the regression coefficients from the robust regression analysis conducted for the 2013 FOEI. The constant reflects the predicted standardised school performance score if the three omitted parent variables were equal to 100 per cent (and all other variables therefore 0 per cent). As the omitted variables are the highest categories for each parental background question, the constant is a relatively high positive value as expected.

The other coefficients represent the degree to which each parental variable increases or decreases predicted school performance relative to the omitted (highest) parent variables. As expected, the majority of coefficients are negative, indicating that the larger the percentage of parents in these categories the lower the predicted level of performance, and hence the higher the level of socio-economic disadvantage.

Only two coefficients are not negative, however both are non-significant in the model (p-values >0.05). These variables are the percentage of parents whose highest level of school education is Year 11 and the percentage of parents whose occupation category is “Other business managers, arts/media/ sportspersons and associate professionals”. Both of these variables are the 2nd highest categories from their respective parental background question, and the coefficients are very small (0.03 and 0.00) indicating that there is no predicted decline in school performance associated with these levels of school education and occupation, relative to the omitted (highest) parent variables.

Table 19:

Robust regression estimated coefficients and statistics

Note: the base or reference category for each parental background item is the highest (omitted) category (i.e., the percentage of parents completing Year 12; with Bachelor degree qualifications; and in the Senior Manager occupation group)

Parental background	Regression parameters	Coef.	Std. Err.	t	P
	Constant	1.114	0.066	16.85	<0.001
School education	Year 9 or equivalent or below	-0.659	0.120	-5.51	<0.001
	Year 10 or equivalent	-0.300	0.079	-3.81	<0.001
	Year 11 or equivalent	0.025	0.138	0.18	0.854
Educational qualifications	No non-school qualification	-1.274	0.121	-10.49	<0.001
	Certificate I to IV	-0.595	0.111	-5.36	<0.001
	Advanced diploma/Diploma	-0.831	0.151	-5.52	<0.001
Occupation group	Not in paid work in last 12 months	-0.887	0.121	-7.35	<0.001
	Machine operators, hospitality staff, assistants, labourers	-0.475	0.120	-3.95	<0.001
	Tradespeople, clerks, skilled office/sales/service staff	-0.354	0.144	-2.45	0.014
	Other managers, arts/media/sports, associate professionals	0.000	0.132	0.00	1.000

Table 19 also indicates that the variable with the greatest predictive relationship with average school performance is the percentage of parents with no educational qualifications. Further, within each parental background question, the general pattern is that the regression coefficients reflect the expected order of the categories, with lower categories generally having a greater negative relationship with average school performance than higher categories.
Within occupation group, the category with the highest negative impact on average school performance is the percentage of parents who are "not in paid work". While this category is potentially substantively different in meaning to the other categories for some groups of parents, this result supports the intended interpretation of this category (i.e., as the lowest occupation status) for the majority of parents.

5.7 Scaling of FOEI scores

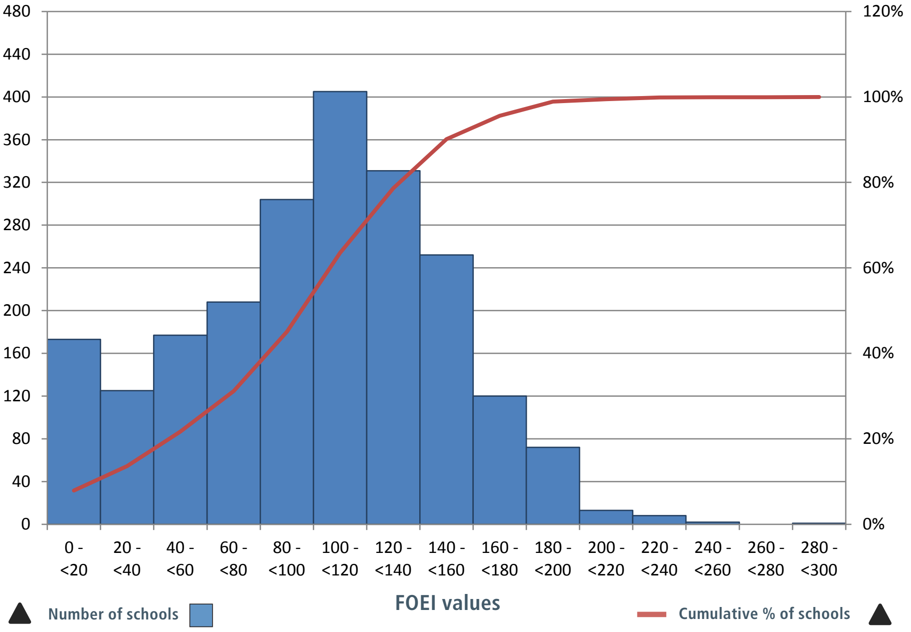
The regression analysis generated predicted scores for all schools[16], whether or not they were able to contribute to the regression model (i.e., including infants schools and senior secondary schools that do not have any NAPLAN results).

The raw predicted scores were initially reversed (by multiplying by -1) so that higher scores correspond to higher levels of socio-economic disadvantage. The scores were then standardised to a mean of 100 and a standard deviation of 50, with a resulting range from -30 to almost 300. Scores below 0 (the most advantaged of schools) were then recoded to 0 for use in the Resource Allocation Model. The distribution of FOEI scores for all NSW government schools is presented in Figure 6.

As shown in Figure 6, FOEI scores are approximately normally distributed. The hump in the distribution for schools with FOEI scores between 0 and 20 reflects the effect of recoding negative values to 0. The most disadvantaged 10 per cent of schools have FOEI scores of 160 or above.

Figure 6:

Distribution of 2013 FOEI scores for all NSW government schools



5.8 Proportion of variance in school performance explained by FOEI

The scatterplots in Figure 7 demonstrate that FOEI is strongly related to average school performance. For both primary and secondary/central schools with more than 100 students (excluding selective schools), FOEI explains almost 80 per cent of the variance in average school performance[17].

[16] This includes only schools that have regular student enrolments. A number of schools such as hospital schools and diagnostic/remedial schools cater to students for very short time periods (e.g., a week or two). FOEI scores have not been calculated for such schools as there is not a relatively stable student population

[17] When all schools are included in the analysis the proportion of variance explained reduces to 63 per cent for primary schools (where average performance is based on at least 3 or more students) and 73 per cent for all secondary/central schools.

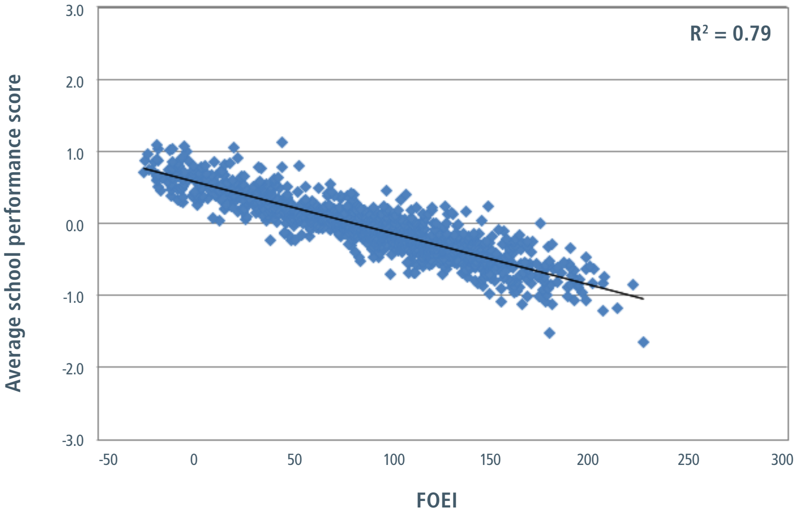
Figure 7:

Scatterplots of 2013 FOEI and 2012 average school performance

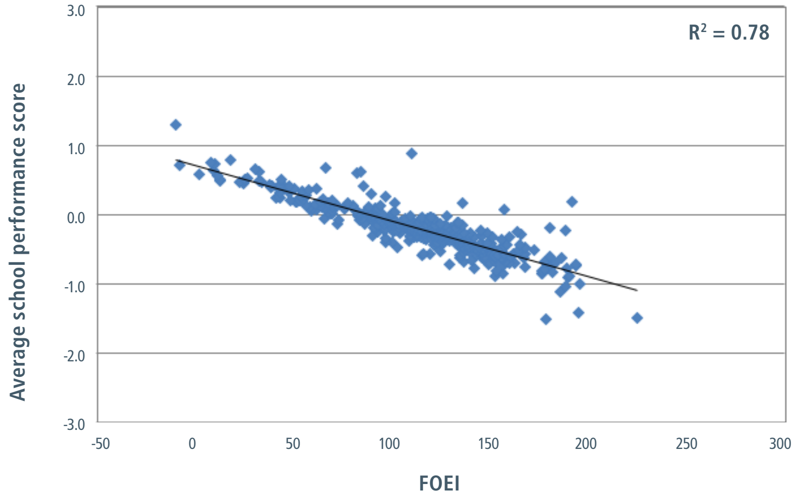
Note 1: This analysis excludes selective schools and schools with fewer than 100 students.

Note 2: The FOEI calculation initially produces some negative FOEI values which are preserved in the graphs above for analysis purposes. When used in RAM, negative values are converted to a value of 0.

(a) Primary schools



(b) Secondary and central schools



5.9 Stability of FOEI over time

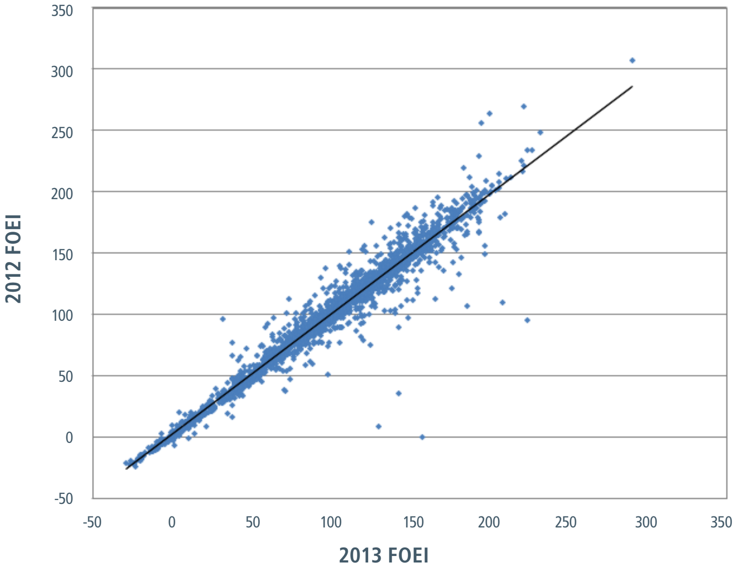
To investigate how stable FOEI scores are from year to year, the FOEI methodology (i.e., multiple imputation for missing data followed by robust regression) was applied to the 2012 student dataset.

As shown in Figure 8, FOEI scores for 2013 are closely aligned with scores for 2012 for most schools. The correlation between the 2 sets of scores is very high at 0.95. For 84 per cent of schools the FOEI scores for the two years are within 10 points, and for 95 per cent of schools the two scores are within 20 points. Almost all the schools where the difference in FOEI scores is greater than 20 points are small schools with fewer than 100 students. Although this pattern is not unexpected (i.e., small schools' FOEI scores are more subject to changes in families' backgrounds than larger schools), the issue of stability of FOEI for small schools is an area for further investigation.

Figure 8:

Scatterplot of 2012 and 2013 FOEI scores

Note: The FOEI calculation initially produces some negative FOEI values which are preserved in the graphs above for analysis purposes. When used in RAM, negative values are converted to a value of 0.



6 Calculation of FOEI quarters

FOEI scores represent the average socio-educational disadvantage of a school's student population. Another important aspect is the distribution of socio-educational disadvantage across the students at the school, as this information is part of the formula to calculate the equity loading for schools. FOEI uses a quarter distribution methodology to represent the student distribution.

To generate each school's quarter distribution:

1. A score was estimated for each student from a combination of parent education and occupation information for students with completed data (either observed or imputed), for each of the 10 'completed' datasets.
    a. The categories for each parent variable were coded from 1 to 4 (or 5 for occupation[18]) as described in Section 3.
    b. For each student, the codes for each of the three parental variables were summed for each parent, and then averaged for students with information for two parents. This resulted in a score for each student that ranged between 3 and 13, with increments of 0.5.
2. From the DEC-wide distribution of students' scores, the cut-points for the 25th, 50th and 75th percentiles were located and used to assign each student to a quarter[19], for each of the 10 'completed' datasets. The resulting cut-points were identical for each of the 10 datasets (at 6.5, 8.5 and 10.5) attesting to the consistency of the imputed data across the 10 imputations.
3. For each school, the percentage of students in each quarter was then determined, for each of the 10 'completed' datasets.
4. The percentage of students in each quarter was then averaged across the 10 'completed' datasets to generate the reported quarter distribution.
5. The number of students in each quarter was then estimated from the reported percentages applied to the total school enrolment[20].

[18] The treatment of occupation as an ordinal variable for the student-level scores is justified by the robust regression results (section 5.6) showing that, for the majority of the parents, "not in paid work" can be considered as the lowest occupation status of all occupation categories.

[19] Due to the distributional properties of the student scores, the numbers of students in each quarter are not exactly equal. The final percentages of students in each quarter, from Quarter 1 to Quarter 4, were 24.5 per cent, 26.6 per cent, 23.1 per cent, and 25.8 per cent.

[20] While a score could not be calculated for students who still had missing parental data even after the imputation process (due to missing data in other variables), these students are included in the calculation of the number of students in each quarter, as the quarter percentages for each school are applied to the school's student population. This effectively assigns students with missing scores to quarters in proportion to the quarter distribution for students with scores.

6.1 2013 FOEI quarter distribution

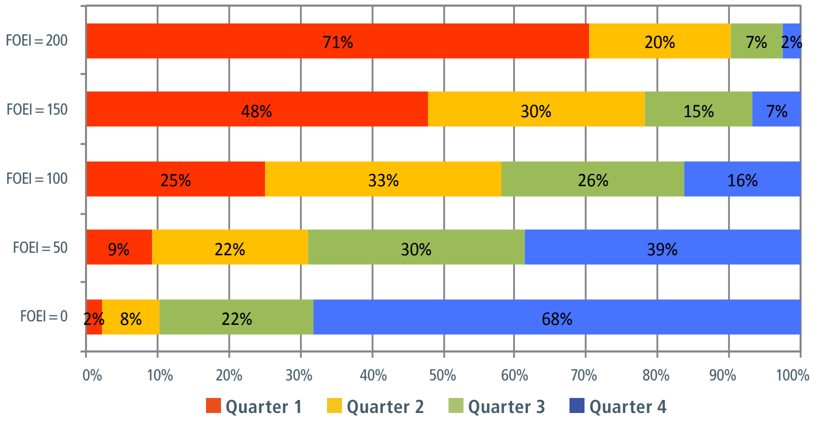
6.1.1 Average quarter profiles for schools with different FOEI scores

Figure 9 shows the average quarter distribution for schools with different FOEI scores. Schools with low FOEI scores (i.e., relatively advantaged) have the majority of their students in Quarters 3 and 4 (around 90 per cent of students on average for schools with FOEI scores close to 0). In contrast, schools with high FOEI scores (i.e., relatively disadvantaged) have the majority of their students in Quarters 1 and 2 (around 91 per cent of students on average for schools with FOEI scores close to 200).

Figure 9:

Average proportions of students in each quarter for schools with different FOEI scores

Note: average quarter proportions are based on the unweighted average of school quarter distributions for all schools with FOEI scores within 5 points of the scores charted



6.1.2 Alignment of FOEI scores and quarters

It is important for the validity and stakeholder acceptance of FOEI scores and quarters that they align well at the school level. That is, a school with a higher FOEI score (i.e., more disadvantaged) should have a greater proportion of its students in Quarters 1 and 2 and fewer in Quarters 3 and 4, than a school with a lower FOEI score. To examine this alignment, quarter distributions for each school have been combined into a composite measure by summing each quarter multiplied by the following factors: Q1 by -1.5, Q2 by -0.5, Q3 by 0.5 and Q4 by 1.5. This composite measure can then be plotted against the FOEI score for each school as shown in Figure 10.

Figure 10:

Alignment of 2013 FOEI scores and quarters

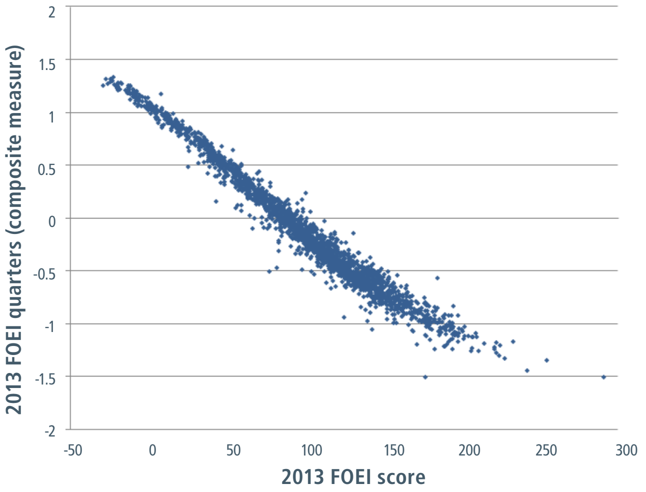


Figure 10 demonstrates that FOEI quarters and FOEI scores align very well. The correlation between FOEI scores and the composite measure for FOEI quarters is very high at 0.98.

The few schools where FOEI scores and quarters are not as closely aligned are very small schools, where a single student can represent a significant percentage of the school population. The student’s location in one quarter or another, particularly when their score is close to the quarter cut-off, can have a large effect on the overall quarter distribution for the school.

7 Quality assurance and further development of FOEI

FOEI will be subject to an ongoing process of review and continuous improvement to ensure that it is the fairest and most robust measure possible of the relative socio-economic disadvantage of NSW government schools. A number of issues have already been identified for further work, as discussed below.

7.1 Improving data quality

A number of strategies have already been implemented to improve the completeness and accuracy of parental education and occupation information recorded in ERN, including:

- Targeted memoranda to school principals informing them of the uses of parental data, the dates when data is extracted for FOEI, and encouraging them to establish processes to periodically review and update parental data in ERN[21].
- The provision of new reports on missing parental data that provide schools with summary information and detailed lists of students with missing data.
- The development of support material for school staff to assist parents select the most appropriate occupation group.
- Modifications to the student enrolment form to elicit higher response rates from parents.

Preliminary analysis of 2014 parental data indicates a significant improvement in the completeness of parental background information since the introduction of FOEI for funding. Information on parental school education is now complete for 94 per cent of parents, and is approaching 90 per cent for parental educational qualifications and occupation group.

Additional work planned includes the development of further support material for schools to assist them to follow up missing data with different parent communities, and liaison with Audit Directorate to ensure that changes to parental background information are included in audit procedures in schools.

7.2 The FOEI construct

The parental background data used for FOEI does not include a component related to family income/wealth or possessions, which is the third main aspect of socio-economic background for children and young people referred to in the literature. Although this information is not suitable for obtaining via enrolment forms, other information that could be used as proxy measures will be explored.

The treatment of the occupation category 'not in paid work' needs further examination, as it could be substantively different in meaning to the other four occupation groups which represent differing skill and status levels. It is likely that this category represents a diverse range of parents, from those who are long-term unemployed through to those wealthier families where one parent can afford to stay home with young children. Possibilities include using this information to calculate an additional variable based on whether neither, one or both parents report being 'not in paid work'.

Additional socio-economic background indicators that will also be investigated for possible inclusion in the FOEI regression model include a single parent indicator, and a measure of student mobility.

[21] Parental background information is collected when a student first enrols in a school and is unlikely to be updated during the time that a student is enrolled in a school unless specific action is taken by the school to review this information with parents. However, for the majority of students, parental information should remain reasonably accurate. The school education level of parents will only change for the very few parents that undertake further secondary-level schooling through TAFE or an equivalent. Parental educational qualifications will only change for the relatively small proportion of parents who complete formal post-school education after enrolling their child at school. Although many parents are likely to change jobs during the time that their children are enrolled in a school they are likely to remain within the same occupation category. The one item which may change is the 'Not in paid work' category. A number of parents re-enter the workforce during the time that their children are enrolled in a school. This is particularly so for mothers who took time out of the workforce to care for young children.

7.3 FOEI methodology

As indicated earlier in this report, a number of methodological issues will be explored in 2014 and future years, including:

- Increasing the number of imputations, as recommended in the recent literature, and analysing the impact of using a larger number of imputations.
- Applying the analysis model to each imputed dataset and then averaging the regression coefficients and predicted values (i.e., FOEI scores) from each dataset. These will be compared to the alternate method used in 2013 to assess any practical differences.
- Using a measurement approach as the basis of the student-level scores. ACARA has recently moved to the use of a measurement model for the student-level measure of socio-educational advantage used to generate ICSEA and its quarter distributions. Alternative methods for constructing the student-level measure for the FOEI quarter distribution will also be further investigated.

7.4 Stability of FOEI

The stability of FOEI for small schools has already been noted as an issue. Further work in this area will explore options to improve the stability of FOEI for small schools without negatively impacting their equity funding. Possibilities include averaging FOEI over two or more years (with possible weighting), or pooling data for students over two or more years.

References

Ainley, J., Graetz, B., Long, M., & Batten, M. (1995). Socioeconomic status and school education. Australian Council for Educational Research.

Allison, P. (2012). Why you probably need more imputations than you think.
from http://www.statisticalhorizons.com/more-imputations (accessed March 26, 2014).

Butler, M. (2012). Coherence of measures of socioeconomic status within and across education and training sectors. Report to the Strategic Cross-sectoral Data Committee, March 2012.

Dong, Y. & Peng, C. J. (2013). Principled missing data methods for researchers. SpringerPlus 2013, 2:222
from http://www.springerplus.com/content/2/1/222 (accessed August 21, 2013).

Graham, J.W. and Olchowski, A.E. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, 8, 206-213.

Kenward, M. & Carpenter, J. (2007). Multiple imputation: current perspectives. Statistical Methods in Medical Research, 16, 199-218.

Lim, P. & Gemici, S. (2011). Measuring the socioeconomic status of Australian youth. National Centre for Vocational Education Research.

Lim, P., Gemici, S., Rice, J. & Karmel, T. (2011). Socioeconomic status and the allocation of government resources in Australia. Education & Training, 53, (7), 570-586.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd edition). New York: John Wiley & Sons.

Marks, G.N. (2011). Issues in the conceptualisation and measurement of socioeconomic background: Do different measures generate different conclusions? Social Indicators Research, 104, 225-251.

Marks, G. N., McMillan, J., Jones, F. L. and Ainley, J. (2000). The Measurement of Socioeconomic Status for the Reporting of Nationally Comparable Outcomes of Schooling. Draft report to the National Education Performance Monitoring Taskforce.

Moons, K.G., Donders, R.A., Stijnen, T. and Harrell, F.E. Jr. (2006). Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology, 59, 1092-1101.

NIASRA (2013). Methodological advice on Family Occupation and Education Index. National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong.

Schafer, J. L. (1999). Multiple imputation: a primer. Statistical Methods in Medical Research, 8, 3-15.

UCLA: Statistical Consulting Group. Stata Annotated Output: Robust Regression.
from http://www.ats.ucla.edu/stat/stata/output/Stata_robust.htm (accessed March 26, 2014).

White, I., Royston, P. and Wood, A. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30, 377-399.

Appendix

Options for combining parent data from 2 parents for each student

There are a number of different ways to combine the information from two parents for each student to generate school-level percentages. The following three methods for constructing the aggregated school-level parent variables were considered in an earlier modelling process:

1. Optimum parent variables – constructed by taking the higher skilled occupation category, the higher school education level and the higher educational qualifications level for each pair of parents.
2. Combined parent variables – constructed by combining the number of first and second parents in each response category.
3. Parent 1 preferred variables – constructed by taking parent 1's response to a question, or if parent 1's response is not available, then parent 2's response, for each pair of parents.

Previous analysis of the relationship between the school-level parental percentages constructed by these 3 different methods, and school performance, indicated very similar levels of association regardless of the method used, as shown in Table A1. Therefore, a decision was made to use the combined parent variables, consistent with the method used for previous versions of FOEI.

Table A1:

Correlations between school-level parental variables constructed by 3 different methods, and average school NAPLAN performance from previous modelling analysis (2012 student enrolments and observed parental data only)

Source: 2012 enrolment data and average school NAPLAN results.

	Response category	Combined Parent	Optimal Parent	Parent 1 Preferred
School Education	Year 9 or equivalent or below	-.705	-.674	-.645
	Year 10 or equivalent	-.562	-.576	-.628
	Year 11 or equivalent	-.416	-.415	-.476
	Year 12 or equivalent	.755	.765	.757
Educational Qualification	Certificate I to IV	-.494	-.494	-.575
	Diploma/Advanced Diploma	.531	.527	.529
	Bachelor degree or above	.751	.750	.738
	No non-school qualification	-.741	-.739	-.700
Occupation Group	Senior management and qualified professionals	.636	.674	.719
	Other business managers, arts/media/ sportspersons and associate professionals	.596	.620	.583
	Tradespeople, clerks and skilled office, sales and service staff	.113	.127	-.250
	Machine operators, hospitality staff, assistants, labourers and related workers	-.476	-.503	-.593
	Not in paid work in last 12 months	-.525	-.549	-.638

Further Information

Statistics Unit
Centre for Education Statistics and Evaluation
NSW Department of Education and Communities
T 02 9561 8691
F 02 9561 8055
www.cese.nsw.gov.au

NSW Department of Education and Communities