An assessment and comparison of methods for non-response adjustment in Australian education surveys

Martin Murphy
University of Wollongong

Follow this and additional works at: https://ro.uow.edu.au/theses1

Recommended Citation
An assessment and comparison of methods for non-response adjustment in Australian education surveys

Martin Murphy

This thesis is presented as part of the requirements for the award of the Degree of Doctor of Philosophy of the University of Wollongong

September 2017

Supervisors:
Prof. David Steel
Dr. Carole Birrell

School of Mathematics and Applied Statistics

This work copyright by Martin Murphy, 2017. All Rights Reserved.

No part of this work may be reproduced, stored in a retrieval system, transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of the author or the University of Wollongong.

This research has been conducted with the support of the Australian Government Research Training Program.
Declaration

I, Martin Murphy declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree Doctor of Philosophy, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Martin Murphy

September 1, 2017
Abstract

This thesis explores methods to address non-response in surveys of Australian school students. These surveys are subject to non-response, resulting in losses in precision and the possibility of non-response bias. Non-response bias is potentially a major source of survey error and protection from this risk is an important consideration in the design of a survey and in the estimation of results. Following a review of the literature of the methods most commonly used to adjust for non-response in surveys, including weighting adjustments, post-stratification, propensity estimation, the use of a generalised regression estimator and regression-based imputation methods, the thesis examines patterns of non-response observed in Australia’s participation in a major international comparative survey, the Trends in International Mathematics and Science (TIMSS) survey conducted in 2011.

The thesis examines factors related to non-response in TIMSS as well as factors that explain mathematics achievement, one of the major outcomes of this survey. A major focus of the research is whether the incorporation of data collected through Australia’s annual census testing of students - the ‘NAPLAN’ assessments – might be used for improving the management of non-response for TIMSS or other such surveys. An additional source of data it will explore for this purpose is MySchool, a web-based resource developed by the Australian Government to assist the parents and other
community members to monitor and compare schools. To date these data sources have not been made available for the management of non-response.

The thesis explores non-response at both school and student levels, reflecting the most commonly used sample designs used in educational surveys and examines and compares adjustment methods across a range of scenarios involving differing rates and complexities of non-response. For non-response at the school level, the widely used practice of substitution of non-responding schools performs well under the scenarios investigated, as do regression based adjustment methods such as the use of a generalised regression estimator and the regression based imputation. For the investigation of student-level non-response, a simulated sub-population of Australian students is constructed using TIMSS survey data.

The investigations in the thesis demonstrate that patterns of non-response are related to important background variables such as the prior performance profile of schools and students and socio-economic background. They also demonstrate that the adjustment methods currently used to address non-response are unlikely to remove non-response bias and may also result in losses of precision. Adjustment methods that make use of NAPLAN and My School data, particularly generalised regression estimation and regression-based imputation methods would likely provide strong protection against non-response bias whilst maintaining good precision on the major survey estimates.
Acknowledgements

This thesis would not have been possible without the support of many people and a number of organisations.

First and foremost I wish to acknowledge the support of my wife Teresa who has shared the burden of this project with me over the many long years it has taken to complete.

I would also like to acknowledge the support of my employer, Australian Council for Educational Research (ACER), and in particular Professor Geoff Masters A.O. and Dr. John Ainley for their support for my endeavours to take ACER’s already strong profile of experience and expertise in large scale survey work, and build on this profile in the area of sampling design and estimation for complex sample surveys. Thank you for the financial support to get me started on this project. Thank you also to Professor Ray Adams who offered his support and expertise early on, and to Dr. Sue Thompson who assisted with my requests for access to data for use in the research. Also special thanks to Ross Turner who was at the front line of handling my various requests for flexibility that were necessary in order to fit this project into my other work responsibilities, and who was always supportive.

It has been a privilege to be able to draw from the extensive knowledge and experience of the staff of the School of Mathematics and Applied Statistics at the University of
Wollongong. In particular, thank you to my supervisor Professor David Steel for his expert knowledge and guidance, from which I have learned so much. Thank you also to my co-supervisor, Dr. Carole Birrell for her contributions, and also other staff for their comments and suggestions early on, including Professor Ray Chambers and Dr. Robert Clark.

The thesis makes use of NAPLAN data provided to me by the Australian Curriculum, Assessment and Reporting Authority (ACARA) and I acknowledge their support in providing me with this data. In accordance with the Deed in Relation to Licence of ACARA Data I insert the following notice:

Notice: Parts of the data used in this publication are sourced from the Australian Curriculum, Assessment and Reporting Authority (ACARA) and are available from ACARA in accordance with its Data Access Protocols.
# Table of contents

## Chapter 1  Non-response in Australian Educational Surveys .................21

1. 1  Introduction ........................................................................................................21
1. 2  Scope of the thesis ............................................................................................... 28
1. 3  Research Questions ............................................................................................. 30
1. 4  Chapter outline .................................................................................................... 33

## Chapter 2  Literature review and formulation of key methods ............... 35

2.1  Introduction ........................................................................................................35
2.2  Non-response as part of Total Survey Error ....................................................... 36
2.3  Accessibility and Amenability ............................................................................. 38
2.4  Managing non-response during design, preparation and data collection ...... 39
2.4.1 Observations from previous surveys ................................................................. 40
2.4.2 Follow-ups during data collection ..................................................................... 42
2.4.3 Two-phased sampling ....................................................................................... 43
2.4.4 The collection of additional data ...................................................................... 45
2.4.5 Substitution ........................................................................................................ 46
2.5  Estimation in the presence of unit non-response ............................................... 47
2.5.1 The potential impact of non-response bias on survey estimation .................. 47
2.5.2 Addressing potential bias through a two-phased sampling design ............... 50
2.5.3 A theoretical framework for non-response management ............................... 52
2.5.4 Incorporating adjustment terms into estimators ............................................ 55
2.5.5 Response probabilities and their correlation with outcomes ....................... 56
2.5.6 Design-based versus model-based approaches to inference ........................ 58
2.6  Non-response adjustment approaches at the estimation stage ....................... 60
2.6.1 Weight classes and post-stratification ............................................................ 60
Chapter 2 

2.6.2 Quasi-randomisation ................................................................. 61
2.6.3 Response homogeneity groups .................................................. 62
2.6.4 Extended probability sampling approach .................................... 63
2.6.5 Imputation ................................................................................. 64
2.6.6 Imputation models .................................................................... 70
2.6.7 Weighting approaches versus imputation .................................... 72
2.6.8 Response propensity estimation .................................................. 72
2.6.9 Estimators directly incorporating auxiliary information – generalised regression estimators .................................................................................. 74
2.6.10 Multiplicative weighting ............................................................. 75
2.6.11 Weight calibration .................................................................... 76
2.6.12 Selecting calibration variables ................................................... 77
2.6.13 Weight trimming ....................................................................... 77
2.7 Formulation of key methods used in non-response ......................... 78
2.7.1 Additional variance under quasi-randomisation ............................ 79
2.7.2 Weighting classes and post-stratification .................................... 81
2.7.3 Modelling of individual response propensities ............................. 84
2.7.4 The Generalised Regression Estimator ......................................... 85
2.7.5 Calibration .................................................................................. 88
2.8 Reporting academic outcomes in educational surveys .................... 89
2.9 Generalised regression estimation in an educational survey .............. 91
2.10 Response propensity adjustments in a study of response bias in England .............................................................................. 92
2.11 Study comparing properties of estimators ....................................... 93
2.12 Chapter Summary ......................................................................... 94

Chapter 3 

Explaining variation in educational achievement ......................... 97
3.1 Introduction ................................................................................... 97
3.2 Australian student surveys – a recent history and background .......... 99
3.2.1 NAPLAN and MySchool ......................................................... 102
Chapter 4  The TIMSS survey and factors explaining achievement........117

4.1  Introduction .......................................................................................... 117

4.2  The 2011 TIMSS survey ........................................................................ 120
   4.2.1 Instrumentation .............................................................................. 121
   4.2.2 Sample Design ............................................................................. 122

4.3  School and Student weight calculations for TIMSS ............................. 128
   4.3.1 School level weights: ................................................................. 128
   4.3.2 Class weights .............................................................................. 131
   4.3.3 Student Weights ......................................................................... 133
   4.3.4 Non-response in additional background instrumentation and data ... 134
   4.3.5 Dependent variable ..................................................................... 135
   4.3.6 Variance estimation .................................................................... 135

4.4  Mathematics achievement outcomes ................................................... 136
   4.4.1 Outcomes from the Australian national report ............................... 136
   4.4.2 Further sub-population comparisons ........................................... 137
   4.4.3 Multiple regression analyses ....................................................... 140
   4.4.4 Sampling Frame variables ......................................................... 143
   4.4.5 Addition of MySchool (school level) variables ............................ 145
   4.4.6 Addition of student Level variables ........................................... 149
4.4.7  Summary of multiple regression analyses of mathematics achievement at the school level ................................................................. 162

Chapter 5  Patterns of response ................................................................. 167

5.1  Identifying absent students ................................................................. 168
5.2  Defining subpopulation categories ....................................................... 170
5.3  Calculating participation rates .............................................................. 171
5.4  Participation rates across population sub-domains ................................ 173
      5.4.1  Socio-economic background measures ...................................... 174
      5.4.2  Performance measures .............................................................. 175
      5.4.3  Participation against performance crossed with socio-economic profile .... 176
      5.4.4  School factors (1) – proportion of ATSI and LBOTE students .......... 178
      5.4.5  School factors (2) – school size; attendance rate .......................... 178
      5.4.6  School factors (3) – school sex composition, school location .......... 179
      5.4.7  State and Sector ........................................................................ 180
5.5  Summary ............................................................................................ 182

Chapter 6  Weighting adjustments in the presence of school non-response 183

6.1  Introduction ....................................................................................... 183
6.2  Approach to the investigation ............................................................. 186
6.3  School level file preparation ................................................................. 187
6.4  The outcome measures to be estimated .............................................. 189
6.5  A school level measure of prior performance ....................................... 189
6.6  Explaining achievement using school level data .................................. 192
6.7  Non-response scenarios and models .................................................... 193
       6.7.1  Missing completely at random non-response ................................ 194
       6.7.2  Missing at random non-response ............................................... 194
       6.7.3  Non-response related to outcomes ............................................ 200
6.8  Expected values of response rates across prior performance and IEO .... 201
6.9  Non-response management approaches ............................................. 205
6.9.1 No school adjustment................................................................. 205
6.9.2 Adjustment by weighting classes within the stratum...................... 206
6.9.3 Response propensity estimation weight........................................... 206
6.9.4 Adjustment by weighting classes defined by response propensity ...... 207
6.9.5 A regression estimator using the 2008 Numeracy mean and the IEO score .... 207
6.9.6 Regression based imputation ......................................................... 207
6.9.7 School substitution........................................................................ 208

6.10 Presentation of outcomes.................................................................. 208
6.11 Bias and precision statistics................................................................. 210
6.12 Selected results.................................................................................. 211
6.12.1 Introduction..................................................................................... 211
6.12.2 MCAR, low response, estimating the mathematics mean ................. 212
6.12.3 MCAR, very low response, estimating the proportion of top students .... 214
6.12.4 MAR1, very low response estimating the mathematics mean .............. 216
6.12.5 MAR 2, low response estimating the proportion in the top Numeracy band...... 219
6.12.6 NMAR, moderate response estimating the Numeracy mean .............. 220
6.12.7 NMAR, very low response estimating the proportion in the top Numeracy band222

6.13 Outcomes summaries........................................................................ 224
6.13.1 Distributions of means, precision and bias ........................................ 224
6.13.3 Distributions of proportions, precision and bias ................................ 227
6.14 Summary of outcomes to approaches for managing non-response .......... 231

Chapter 7 Constructing a simulated student population ............................. 234
7.1 Introduction ....................................................................................... 234
7.2 Constructing a student level population file ......................................... 238
7.3 Imputation algorithm ......................................................................... 239
7.4 Records available for imputation........................................................ 242
7.5 Imputation outcomes at the school level .............................................. 243
Chapter 8  Estimation of student outcomes under non-response........258

8.1 Introduction .................................................................................. 258
8.2 Approach to the investigation............................................................. 260
8.3 School and within-school design weights............................................ 262
8.4 Key variables used in the analysis...................................................... 263
8.5 The population values to be estimated............................................... 264
8.6 Models explaining variation in outcomes........................................... 266
8.7 Non-response scenarios investigated................................................ 267
  8.7.1 Inducing missing completely at random (MCAR) non-response ........ 269
  8.7.2 Inducing missing at random (MAR) non-response......................... 270
  8.7.3 Inducing non-response not explained by the survey design variables (NMAR). 273
8.8 Adjustment methods ........................................................................ 277
  8.8.1 No non-response adjustment......................................................... 277
  8.8.2 Method 1: A school level non-response adjustment weight equal to the number who were sampled divided by the number who responded ................................................ 277
  8.8.3 Method 2: non-response adjustment by sex within school, assuming at least five responding boys and girls. Otherwise method 1................................................ 278
  8.8.4 Method 3: Post-stratification adjustments to design weights by n(samp)/n(resp) within classes defined across the population by prior performance home educational resources and sex.......................................................... 279
  8.8.5 Method 4: estimating the response propensity using logistic regression, and adjusting the design weight by the inverse of the estimated propensity. ........... 279
  8.8.6 Method 5: non-response adjustments to the design weights within weighting classes defined by quintiles of the response propensities estimated in method 4. 280
  8.8.7 Method 6: Using a generalised regression estimator based on the relationship between maths score and prior performance, home educational resources with estimates weighted by the design weight.................................................. 280
8.8.8 Method 7: An imputation of the maths outcomes using a regression model based to predict mathematics achievement by science performance, HER and sex, on the respondent data for each sample. .................................................. 281

8.8.9 Method 8: multiple imputation of mathematics outcomes on prior performance, home educational resources and sex. .................................................. 281

8.9 Approach to presenting outcomes .................................................. 282

8.9.1 Overall approach ............................................................................. 282

8.9.2 Graphical and tabular representation of selected outcomes ................. 282

8.10 Selected results by scenario and method ............................................ 284

8.10.1 Estimating the mathematics mean - MCAR1 .................................. 284

8.10.2 Estimating the proportion with scores over 430 in Maths – MCAR1 .......... 286

8.10.3 Estimating the mathematics mean - MCAR2 .................................. 287

8.10.4 Estimating the proportion with scores over 430 in Maths – MAR1 .......... 289

8.10.5 Estimating the mathematics mean – MAR2 .................................. 291

8.10.6 Estimating the mathematics mean – NMAR1 .................................. 292

8.10.7 Proportion with scores over 430 in Maths - NMAR1 ......................... 294

8.10.8 Weight diagnostics: NMAR1 .................................................. 295

8.10.9 Estimating the mathematics mean – NMAR3 .................................. 296

8.10.10 Weight diagnostics: NMAR3 .................................................. 297

8.11 Results summaries ............................................................................. 298

8.11.1 Shape, precision and bias: Estimating the population mean - all scenarios...... 298

8.11.2 Shape, precision and bias: Estimating the population proportion - all scenarios. 302

8.12 Summary of findings of student non-response investigations .................. 306

Chapter 9 Summary and conclusions .................................................. 311

9.1 Non-response and the quality of survey outcomes .................................. 311

9.2 Minimising non-response during the data collection phase ....................... 312

9.3 The approaches used to manage non-response ...................................... 313

9.4 Understanding the factors relating to achievement .................................. 313

9.5 Understanding the factors driving non-response ...................................... 314
9.6 Non-response in a two-stage sample design................................................................. 315

9.6.1 School level non-response ......................................................................................... 315

9.6.2 A database for exploring student level non-response ............................................. 319

9.6.3 Student level non-response ..................................................................................... 320

9.7 Conclusions and further investigations...................................................................... 323

9.7.1 Recommendations for future Australian educational surveys .............................. 325

9.7.2 Technical issues for further investigation ................................................................. 327

References 331

Appendix A: Diagnostics of multiple regression explaining TIMSS mathematics achievement 339
List of Tables

Table 3-1: National Assessment Program surveys 2011-2017 .............................................. 101
Table 4-1: Exclusion rates TIMSS 2011 Year 8 Australia ....................................................... 123
Table 4-2: School participation by State and Territory, TIMSS 2011 Year 8 ............... 126
Table 4-3: Student participation by State and Territory, TIMSS 2011 Year 8 .......... 127
Table 4-4: Year 8 Mathematics outcomes by school sex composition, TIMSS 2011 .... 137
Table 4-5: Year 8 Mathematics outcomes by sector, TIMSS 2011 ....................... 137
Table 4-6 Year 8 Mathematics outcomes by SEIFA IEO, TIMSS 2011 ..................... 138
Table 4-7: Year 8 Mathematics outcomes by average income, TIMSS 2011 ........... 138
Table 4-8: Year 8 Mathematics outcomes by Home Education Resources index, TIMSS 2011 .......................................................... 139
Table 4-9: Year 8 Mathematics outcomes by ICSEA percentile groups, TIMSS 2011139
Table 4-10: Year 8 Mathematics outcomes by school mean numeracy score at Year 9,
TIMSS 2011 .................................................................................................................. 140
Table 4-11: Year 8 Mathematics outcomes by school size groups, TIMSS 2011 .... 140
Table 4-12: Multiple regression on mathematics achievement, TIMSS 2011 Year 8,
Model 1: Sampling Frame variables ................................................................. 144
Table 4-13: Multiple regression on mathematics achievement, TIMSS 2011 Year 8,
Model 2: Addition of MySchool variables...................................................... 147
Table 4-14: Multiple regression on mathematics achievement, TIMSS 2011 Year 8,
Model 3: Addition of student level variables from TIMSS ......................... 152

Table 4-15: Multiple regression on mathematics achievement (extract), TIMSS 2011
Year 8, Models 3a-3e: socio-economic variables ........................................ 153

Table 4-16: Variance Inflation Statistics on socio-economic variables by regression
model........................................................................................................ 154

Table 4-17: Multiple regression on mathematics achievement, TIMSS 2011 Year 8,
Model 4: addition of prior achievement variable ...................................... 156

Table 4-18: Multiple regression on mathematics achievement, TIMSS 2011 Year 8,
Model 4a: removal of school NAPLAN Numeracy mean ......................... 158

Table 4-19: Developed scales based on student, school and teacher questionnaires:
TIMSS 2011 ................................................................................................. 159

Table 4-20: Multiple regression on mathematics achievement, TIMSS 2011 Year 8,
Model 5: addition of school climate and student affect variables .......... 161

Table 4-21: Multiple regression on mathematics achievement, TIMSS 2011 Year 8,
model summary.......................................................................................... 163

Table 5-1: School and student participation, TIMSS 2011 Year 8, Australia .......... 168

Table 6-1: Multiple regression on mathematics achievement, TIMSS 2011 Year 8.
Comparing performance measures.............................................................. 191

Table 6-2: Multiple regression on mathematics achievement, TIMSS 2011 Year 8.
Comparing socio-economic background measures..................................... 193
Table 6-3: Expected value and standard error of Numeracy mean under MCAR school non-response with 18 respondents ..........................................................213

Table 6-4: Expected value and standard error of proportion in top Numeracy band under MCAR school non-response with 9 respondents ........................................215

Table 6-5: Expected value and standard error of Numeracy mean under MAR (v1) school non-response with 9 respondents ..................................................217

Table 6-6: Expected value and standard error of proportion in top Numeracy band under MAR (v2) school non-response with 18 respondents .........................220

Table 6-7: Expected value and standard error of Numeracy mean under NMAR school non-response with 27 respondents ....................................................221

Table 6-8: Expected value and standard error of proportion in top Numeracy band under NMAR school non-response with 9 respondents ..........................223

Table 6-9: Precision and bias. Estimates of population mean. All non-response models ..........................................................226

Table 6-10: Precision and bias. Estimates of proportion. All non-response models ....230

Table 7-1: Imputations summary – Victorian government schools .........................243

Table 7-2: Comparison of original records and imputed records: TIMSS 2011 Science Year 8 Mathematics .................................................................................245

Table 7-3: Comparison of original records and imputed records: TIMSS 2011 Science Year 8 .........................................................................................248

Table 7-4: Comparison of original records and imputed records: TIMSS 2011 Year 8: Student Age .........................................................................................249
Table 7-5: Comparison of original records and imputed records: TIMSS 2011 Year 8
Home Educational Resources.................................................................250

Table 7-6: Intra-class correlations of Maths, Science, Home Educational Resources –
original versus imputed.................................................................254

Table 7-7: Intra-class correlations with additional clustering factor added to imputation
model.................................................................................................256

Table 7-8: Frequencies of class sizes within schools following imputation.............257

Table 8-1: Design weights by school ID. Simulated population. .....................263

Table 8-2: Regression parameters: Mathematics achievement explained by science
achievement, home educational resources and sex, unweighted ............266

Table 8-3: Regression parameters: Mathematics achievement explained by science
achievement, home educational resources and sex, weighted by school
weight.................................................................................................267

Table 8-4: Mean estimates by non-response management method. MCAR 1........284

Table 8-5: Estimated proportions by non-response adjustment MCARI ............286

Table 8-6: Weight diagnostics: Estimates of population mean - MCAR 1............287

Table 8-7: Mean estimates by non-response adjustment. MCAR 2....................288

Table 8-8: Weight diagnostics: Estimates of population mean - MCAR2...........289

Table 8-9: Estimated proportions by non-response adjustment MAR1...............289

Table 8-10: Weight diagnostics: Estimates of population proportion – MAR1 ....291

Table 8-11: Mean estimates by non-response adjustment. MAR2 ....................291
Table 8-12: Mean estimates by non-response adjustment. NMAR1 ..............................293

Table 8-13: Proportion estimates by non-response adjustment. NMAR1 ......................294

Table 8-14: Weight diagnostics: Estimates of population mean - NMAR 1 ..................295

Table 8-15: Mean estimates by non-response management method. NMAR3 ............296

Table 8-16: Weight diagnostics: Estimates of population mean – NMAR3.................297

Table 8-17: Precision and bias statistics – estimated mean under student level non-
response ..................................................................................................................300

Table 8-18: Precision and bias – estimated proportion under student level non-response
........................................................................................................................................305
List of Figures

Figure 4-1: Student Questionnaire items used in the Year 8 Home Educational Resources Scale, TIMSS 2011 ................................................................. 150

Figure 5-1: Participation rates by ICSEA and HER, TIMSS 2011 Year 8, Australia .. 174

Figure 5-2: Participation rates by SEIFA IEO, TIMSS 2011 Year 8, Australia........... 175

Figure 5-3: Participation rates by performance measures, TIMSS 2011 Year 8, Australia ........................................................................................................ 176

Figure 5-4: Participation rates by performance x socio-economic background, TIMSS 2011 Year 8, Australia ........................................................................ 177

Figure 5-5: Participation rates by ATSI, LBOTE proportions, TIMSS 2011 Year 8, Australia ........................................................................................................ 178

Figure 5-6: Participation rates by school size and attendance rate, TIMSS 2011 Year 8, Australia ........................................................................................................ 179

Figure 5-7: Participation rates by school sex composition, TIMSS 2011 Year 8, Australia ........................................................................................................ 180

Figure 5-8: Participation rates by location, TIMSS 2011 Year 8, Australia............. 180

Figure 5-9: Participation rates by State and Territory, TIMSS 2011 Year 8, Australia 181

Figure 5-10: Participation rates by sector, TIMSS 2011 Year 8, Australia ............... 181

Figure 6-1: Measures of size (response probability) by 2008 Numeracy mean x IEO quintile: School non-response model ‘MAR1’ ............................................. 197
Figure 6-2: Measures of size (response probability) by 2008 Numeracy mean x IEO quintile: School non-response model ‘MAR2’ .............................................. 199

Figure 6-3: Measures of size (response probability) by 2008 Numeracy mean x IEO quintile: School non-response model ‘NMAR’ ................................................ 201

Figure 6-4: Expected value of the response rate used in the MCAR non-response model across performance x socio-economic background categories .................. 203

Figure 6-5: Expected value of the response rate used in the MAR1 non-response model across performance x socio-economic background categories .......... 204

Figure 6-6: Expected value of the response rate used in the MAR2 non-response model across performance x socio-economic background categories .......... 204

Figure 6-7: Expected value of the response rate used in the NMAR non-response model across performance x socio-economic background categories .......... 205

Figure 6-8: Box plot description from (SAS/GRAPH(R) 9.2: Statistical Graphics Procedures Guide). ........................................................................................................ 209

Figure 6-9: Distribution of estimated mathematics mean under MCAR school non-response with 18 respondents ................................................................. 213

Figure 6-10: Distribution of estimated proportion in top Numeracy band under MCAR school non-response with 9 respondents ........................................ 215

Figure 6-11: Distribution of estimated mathematics mean under MAR (v1) school non-response with 9 respondents ................................................................. 217

Figure 6-12: Distribution of estimated proportion in top Numeracy band under MAR (v2) school non-response with 18 respondents ........................................ 219
Figure 6-13: Distribution of estimated mathematics mean under NMAR school non-response with 27 respondents ................................................................. 221

Figure 6-14: Distribution of estimated proportion in top Numeracy band under NMAR school non-response with 9 respondents ......................................................... 223

Figure 6-15: Summary distributions – estimating the population mean – school level non-response .................................................................................................. 225

Figure 6-16: Summary distributions – estimating the population proportion – school level non-response .................................................................................................. 229

Figure 7-1: TIMSS mean versus NAPLAN mean – Victorian government versus other jurisdictions ........................................................................................................ 240

Figure 7-2: Scatterplot of TIMSS mean versus NAPLAN mean – Near neighbours ... 241

Figure 7-3: TIMSS 2011 Y8 Mathematics mean scores – original versus combined records .............................................................................................................. 246

Figure 7-4: TIMSS 2011 Y8 Standard deviation of Mathematics scores – original versus combined .............................................................................................................. 247

Figure 7-5: TIMSS 2011 Y8 Science mean scores – original versus combined ............... 251

Figure 7-6: TIMSS 2011 Y8 Standard deviation of Science scores – original versus combined .............................................................................................................. 252

Figure 7-7: TIMSS 2011 Y8 Home Educational Resources mean scores – original versus imputed .......................................................................................................... 252

Figure 7-8: TIMSS 2011 Y8 Standard deviation of Home Educational Resources – original versus combined ................................................................. 253
Figure 8-1: Assumed response rates: Student non-response MAR (v1) model ..........271
Figure 8-2: Assumed response rates: Student non-response MAR (v2) model ..........273
Figure 8-3: Assumed response rates: Student non-response NMAR (v1) model ........274
Figure 8-4: Assumed response rates: Student non-response MAR (v2) model ..........275
Figure 8-5: Assumed response rates: Student non-response NMAR (v3) model .......276
Figure 8-6: Distribution of mean estimates by non-response adjustment - MCAR1....285
Figure 8-7: Distribution of proportion estimates by non-response adjustment - MCAR1
..................................................................................................................................286
Figure 8-8: Distribution of mean estimates by non-response adjustment MCAR2.....288
Figure 8-9: Distribution of proportion estimates by non-response adjustment – MAR1
..................................................................................................................................290
Figure 8-10: Distribution of mean estimates by non-response adjustment MAR2......292
Figure 8-11: Distribution of mean estimates by non-response adjustment NMAR1....293
Figure 8-12: Distribution of proportion estimates by non-response adjustment – NMAR1...........................................................................................................295
Figure 8-13: Distribution of mean estimates by non-response adjustment NMAR3....297
Figure 8-14: Summary distributions – estimating the population mean – student level non-response........................................................................................................299
Figure 8-15: Summary distributions – estimating the population proportion – student level non-response........................................................................................................304
Chapter 1 Non-response in Australian Educational Surveys

1.1 Introduction

This thesis explores methods that can be used to address non-response in surveys of Australian school students. As with surveys in most fields, Australian educational surveys experience non-response and are therefore subject to losses in accuracy and precision. With respect to the accuracy of estimates from surveys subject to non-response, there is the issue of potential non-response bias, the possibility that sampled respondents differ from the non-respondents on outcomes of the survey, leading to biased estimates. Survey bias is potentially a major source of survey error and protection from this risk is an important consideration both with respect to the survey design, as well as in the selection of methods used for the estimation and analysis of collected data. With respect to survey precision, the loss of data through non-response will generally increase the size of standard errors and confidence intervals of major survey outcomes. This occurs directly as a result of the smaller achieved sample size, and also indirectly through design effects associated with for example the increased variation in weights designed to address non-response.
Since the 1990’s Australian policy makers have specifically pursued a policy of “using data collections to improve Australian education policy” (Ministerial Council On Education Employment Training And Youth Affairs Performance Measurement and Review Taskforce, 2006) and have undertaken surveys to obtain estimates across a range of aspects of student development – curriculum, attendance, and generic skills and attributes. A detailed summary of educational survey activity over the last decade will be presented in Chapter 3 of this thesis. The outcomes from these surveys have become a very important source of information used by policy makers to guide policy and to inform the distribution of resources towards policy priorities.

Survey data is also made available more widely to drive improvements at the local level. Extensive data are published on the MySchool website which provides students and parents with information on each school and publishes comparisons between schools. ‘These comparisons provide information to support improvements in schools’ (ACARA (2017b)). MySchool will be described more fully in Chapter 3.

The increased focus on survey-based analyses of educational outcomes in Australia is paralleled with similar activities internationally. Educational achievement indices are increasingly viewed as key indicators of the prosperity, and future economic potential of countries. “The prosperity of countries now derives to a large extent from their human capital, and to succeed in a rapidly changing world, individuals need to advance their knowledge and skills throughout their lives. Education systems need to lay strong
foundations for this, by fostering knowledge and skills and strengthening the capacity and motivation of young adults.” (OECD, 2004)

In addition to the conduct and organisation of national surveys, Australia has participated in a number of major international comparative educational surveys. These have included a number of surveys conducted by the International Association for the Evaluation of Educational Achievement (IEA) - for example the study investigating Trends in Mathematics and Science achievement (TIMSS) and the Progress in International Reading Literacy (PIRLS) study - and also the Programme of International Student Assessment (PISA), organised by the Organisation for Economic Cooperation and Development (OECD).

The delivery of a successful education system is a complex and costly endeavour, and there is considerable interest among countries in identifying successful educational systems internationally, and the factors underlying these successes. A great deal of effort is therefore invested in the design of surveys to ensure that useful and important international comparisons are achieved. Considerable input is provided from experts in the fields of education specifically as well as in experts in the design of large scale surveys. The aforementioned surveys – TIMSS, PIRLS and PISA - are generally very highly regarded for the quality of their implementation. They have been influential in the selection of methods and practices for Australian surveys, including methods in relation to the management of non-response.
‘Managing non-response’ generally involves drawing on data known about the missing schools and students, or collected from the survey or from other sources to make reasonable estimates of those missing data. Methods include giving responding schools and students considered similar to those missing school and students a larger influence on outcomes by increasing their weight. Also it may involve making imputations, or predictions for those missing data through observations made using the respondent data about the relationships between variables and outcomes. Approaches to the management of non-response currently used in these surveys are discussed in detail in Chapter 2.

The methods used in the international surveys have been subject to some debate and discussion. In the early 2000’s a lively debate was generated following the rather dramatic exclusion of United Kingdom’s (UK’s) data from the international database and comparison tables for the 2003 PISA survey as a result of a failure to meet the minimum participation rate standards set for this survey – briefly 85% of sampled schools, and 80% of students from within those schools. Prais (2003) questioned a number of features of the approach to estimation in the PISA survey, including a strong critique of the manner in which ‘replacement’ schools were used as substitutes for sampled but non-responding schools.

An interesting component of the discussion arising from the UK’s exclusion from the PISA comparison tables was the fact that the UK educational system has very detailed information about the sampled students who did not participate in the survey and their schools, collected through their national assessment regime. As will be discussed in the review of methods for managing non-response in Chapter 2 (section 2.10), the UK
government commissioned a study Micklewright, Schnepf, and Skinner (2012) to investigate whether these national data could have been used to address the concerns about non-response bias arising from the lower response rates.

In recent years, Australia has moved in a similar direction with respect to the collection and organisation at a national level of very detailed information about schools and their students, particularly through the activities of the National Assessment Program (NAP), which will be described in Chapter 3.

While the methods of the international comparative education surveys are highly regarded, the data used in the management of non-response and in analyses for these surveys tend to be limited to data collected from the survey itself, or sample design data – data provided on national databases used in sample selection. (School sample selection is almost always done by the international centre). There are good reasons why national data outside of that provided on the sampling frame are very difficult to incorporate into non-response management methods and analyses for these surveys. A key principle of these surveys is comparability across all dimensions of the survey activity, including in the methods used for managing non-response. The incorporation of data external to the survey would entail another whole layer of quality monitoring of those other data. With more than 70 participating countries for some of these international surveys, it is beyond their scope to evaluate the relative quality of data collected through separate survey activity within individual countries.
Furthermore the primary outcome objectives for international surveys are estimates at the national level, and the methodologies for these surveys are designed to be optimised for those outcomes. The methodologies used may therefore not address local, sub-national diversities as fully as surveys designed specifically for a national context may be able to. The Australian States and Territories are primarily responsible for public school education, and there is additionally, a substantial non-government education sector, and so outcomes and comparisons at the levels of State and/or Sector are of interest within the Australian context. It may be the case that non-response adjustment methods that work best for subpopulations such as State and/or Sector differ to those that are deemed adequate for the national level comparisons desired for the international studies.

The NAPLAN census assessments themselves experience non-response. As always with the incidence of non-response there is the concern that it may be related to the performance outcomes being measured and estimated, and may therefore lead to biased outcomes. Reports by the Australian Curriculum and Assessment Authority (ACARA), which administers the NAPLAN assessments, have noted small numbers of cases where schools have “attempted to exert influence with regard to” participation in the NAPLAN assessments, for example (ACARA Australian Curriculum Assessment and Reporting Authority, 2011). A 2012 report to the Council of Australian Governments (COAG) includes outcomes from an exploratory study on the relationship between participation and achievement with analysis that “showed that there were differences in student performance based on whether they participated or not. Students who were present had
the highest average scores but absent students had the lowest, based on estimating likely scores from their background characteristics.” (COAG Reform Council, 2012).

A number of articles have been written expressing the concern that students may have been actively encouraged to not participate in these assessments for fear of negative publicity regarding school level performance, see for example Holden (2010). Education Columnist Maralyn Parker, writing for The Daily Telegraph newspaper wrote: “But, as the anti-NAPLAN activists know, the majority of Australian parents don't need to boycott the tests for the national testing agenda to be undermined. All that is needed is for around 10 per cent of children at a school to be withdrawn from the tests for the school's averages to no longer be credible or useful.” Parker (2012).

As with other surveys, methods are applied to NAPLAN data to adjust for the incidence of non-response, but to date, these methods have not made use of much of the data collected during a student’s progression through school.

While the NAP comprises a quite extensive suite of international and national surveys of students covering all Australian jurisdictions, other surveys of students are conducted in Australia each year. These include for example surveys undertaken by individual jurisdictions, a number of health surveys and others. It is of interest to investigate whether the incorporation of data collected through the NAP might lead to supplementary, or altogether different approaches to non-response management for the NAP surveys themselves, as well as these other educational surveys conducted.
This thesis aims to look ahead to a time probably not very far into the future, where a wider set of the data collected about Australian schools and students – e.g. MySchool data, NAPLAN data - will be more accessible to the statistician managing non-response in educational surveys. Could these data be used to better address the management of non-response than the current approaches? Would approaches that incorporate these additional data help to address concerns about non-responding students being withdrawn from assessments? Are there efficiencies or other improvements to be gained from an approach that makes use of these data compared to the current approaches?

1.2 Scope of the thesis

The thesis will focus its investigation on surveys that estimate the academic outcomes of Australian school students. A detailed examination will be conducted of the non-response methods that are used, and their relationship to the outcomes of major national and international educational surveys conducted in Australia in recent years. Most attention will be paid to the details of the TIMSS survey, which has been conducted in Australia over a period of more than twenty years. Where appropriate reference will also be made to methods used in other surveys conducted under Australia’s NAP. Later parts of the thesis will make use of data from TIMSS to investigate and evaluate possible alternative approaches to managing non-response.

The approach to managing non-response requires consideration of the sample design. All of the surveys mentioned above have similar sampling designs – a stratified cluster based approach, with schools sampled with probability proportional to size from a list of all eligible schools, followed by the selection of an approximately equal number of
students from each sampled school, either the selection of an intact class of students, or a selection of a fixed number of students from a list of all eligible students from the school. The thesis will limit its attention to surveys that fall under this broad sampling design.

While the main focus of the thesis will be on large-scale national surveys, it aims to provide a framework for decisions made regarding non-response for surveys of smaller scope. Participation in the NAP assessments is tied to school funding, having the effect of essentially mandating school participation in these surveys. Other surveys of Australian students do not fall under the NAP framework, and are therefore much more likely to experience non-response at the school level. In fact non-response at this level is likely exacerbated by the compulsory nature of participation in the NAP surveys. With substantial school-level non-response, different methods for managing non-response need to be considered. For example, the thesis will explore the issues surrounding the common approach of school substitution as a method for managing non-response in these surveys.

Factors that explain an individual’s performance on an academic assessment are likely very complex, and extend beyond matters such as the student’s school, sex or socio-economic background. It is not the intention of this thesis to attempt to explore factors that might explain educational performance beyond those that might reasonably be imagined as available to the statistician managing survey non-response issues. These will include school level variables such as geographic location, sector and socio-
economic profile, as well as individual characteristics such as sex, age, prior performance and socio-economic background.

1. 3 Research Questions

The primary question of interest for the thesis is:

*What improvements might be expected in the accuracy and precision of estimates from surveys of Australian school students if data collected through Australia’s National Assessment Program was incorporated into non-response management?*

With the clustered sampled designs used for surveys under consideration for this thesis some information is known about all students sampled to participate in the survey, even those who do not participate. If the non-participation arises at the school level, then information about the school is obtainable from either the database of schools used for sampling, or from other sources such as the MySchool website. School level information such as the geographic location, enrolment structure, sectoral classification, as well as measures of average socio-economic level and average performance on prior assessments are obtainable. If the non-participation arises for individual students within schools, then additional information may include the sex, age and year level, as well as individual measures of socio-economic background, and performance on prior assessments. In the PISA survey for example, lists of eligible students are prepared for each sampled, participating school, from which a sample of a fixed number of students is selected. These lists include the age, sex, year level and study program of the student.
Of particular interest is whether the outcomes from the NAPLAN census-based survey as well as data collected annually about schools to describe the composition of the school student intake could be used to contribute to non-response management of other student surveys undertaken in Australia.

This question will be broken into two components, reflecting the two-stage nature of major surveys of students in Australia:

For each stage of sampling:

*What non-response management method appears most successful at managing non-response? What improvements in precision and bias could be expected with these methods if NAP data was incorporated?*

At the estimation stage following data collection, non-response management used for most Australian surveys in recent years has been limited largely to adjustments based on broader school level variables such as the State, sector and geographic location of the school. In a small number of cases, finer adjustments, for example by sex within school, have been considered and/or incorporated. But in general the data collected on student performance through NAPLAN and other factors such as student socio-economic background have not been made available for non-response management. As these data are likely to be well correlated with academic outcomes for a survey, some improvements would be expected with the incorporation of these data into non-response management approaches. The investigation into non-response management at each stage
of sampling – schools, then students - will examine what degree of improvements might be expected if these external data were made available.

Non-response management extends to approaches adopted prior to and during data collection. Of particular note is the use of substitutes for non-responding schools, a practice widely used in Australian surveys and internationally, including in TIMSS, PIRLS and PISA. As noted above, the use of substitutes was subject to some criticism in the discussions arising from the UK’s removal from international comparison tables for the PISA 2003 survey. As part of the investigation into managing school-level non-response, this practice will be examined and compared with other management approaches. None of the aforementioned surveys permit substitutes for non-responding students.

As noted above, the most successful non-response management methods are likely to involve the use of auxiliary data that contribute to explaining survey outcomes, (e.g. achievement) and/or non-response. An important preliminary question which will be addressed in the early chapters of the thesis is therefore the extent to which the data available for non-response management appears to explain achievement and non-response in the Australian context:

*What variables collected through Australian student surveys are most important in explaining achievement and survey non-response?*
1. 4 Chapter outline

Following a review of the statistical literature on methods for managing non-response in Chapter 2, Chapter 3 will review recent studies into the factors explaining academic performance in the Australian context. Following a detailed description of the 2011 TIMSS survey instrumentation and weighting, Chapter 4 includes an investigation of the relationship between a range of available variables at both school and student level and performance using multiple regression. Chapter 5 will extend this investigation of TIMSS 2011 survey data to examine the factors most important in explaining patterns of response. Chapter 6 will draw from the observations of these investigations from Chapters 4 and 5 to examine the efficacy of a range of non-response management approaches under various school level non-response scenarios. As preparation for an investigation of non-response at the student level, Chapter 7 will detail the preparation of a simulated student population data file derived from the TIMSS database. Chapter 8 will use this simulated population to induce non-response at the student level, and to apply and compare a range of methods for addressing this non-response. Chapter 9 will bring the results of the investigations from the previous chapters together to draw conclusions about the effectiveness of the relatively simple non-response management practices currently in place for Australian educational surveys and whether gains in both the precision of estimates of survey outcomes and in the protection from non-response bias indicated from those investigations might justify the adoption of more sophisticated approaches to non-response management into the future.
Chapter 2  Literature review and formulation of key methods

2.1 Introduction

The first part of this chapter summarises the literature on the development of non-response and the development of new methodologies and enhancements in management practices over time. Following this review, key aspects of the formulation of the methodologies used in later chapters of the thesis are illustrated. The chapter concludes with a description of three studies that apply a number of the methods described.

An awareness of the potential impact of non-response on survey estimation has been present ever since probability-based designs became the predominant approach to survey work. As far back as 1946, Hansen and Hurwitz (1946) developed the idea of two-phased sampling as a measure to reduce potential bias arising from non-response.
There is now a substantial literature associated with the issue of non-response, for example Little and Rubin (2014); Särndal and Lundström (2005); Groves, Dillman, Eltinge and Little (2002); Lessler and Kalsbeek (1992); Haziza (2009); Brick and Montaquila (2009). The literature spans all aspects of survey activity, from the preparation for field work; the field work itself, and the estimation and analysis of collected data.

2.2 Non-response as part of Total Survey Error

There are two broad categories of non-response distinguished in the literature, item- and unit- non-response.

*Item non-response* refers to the situation where responses to some but not all data items have been collected from the unit selected to participate in the survey. Some items may be missing or invalid, for example ‘not-reached’ items on a test. In the case of item non-response, the fact that some data have been collected from the sampled unit will normally provide an avenue for addressing the missing item data.

*Unit non-response* refers to the failure to receive any survey data from units sampled to participate in the survey, for example because a student is absent from a test, or because a school declines to participate in the survey. While no information has been provided directly from non-responding units, there will often be other auxiliary information available about these units, for example the location and socio-economic profile of the student’s school, or prior performance of the student on a test. It may be possible to use these auxiliary data to make adjustments to estimates derived from the responding sample to improve the estimation of population characteristics.
The focus of this thesis is on approaches to managing the potential for bias and loss of precision in the context of unit non-response – when whole schools or individual students fail to respond to the survey.

Considerable work has been done in recent years towards establishing a theoretical platform for understanding the factors that lead to non-response, and to bring together the various aspects that contribute to so-called total survey error, including coverage problems, measurement errors, sampling errors, as well as errors due to non-response. Groves and Couper (1998) have attempted to classify various aspects of the data collection process that contribute to survey error.

Within the Total Survey Error framework, errors are classified under two broad classes:

- Errors of observation, which includes coverage and sampling related errors, including errors arising from non-response;
- Measurement errors, including errors with respect to the measurement construct, errors of measurement, data processing errors and so on.

Among the class of observation errors, errors of non-response are distinguished from errors arising from the preparation of the sampling frame (e.g. errors of coverage), and from sample selection (sampling error). Non-response errors arise during the attempt to collect data from the sampled units. Following data collection, a further potential source of error - adjustment error – arises from the attempts to adjust for the non-response that has occurred.
Lessler and Kalsbeek (1992) provide expressions for the components of non-response within the context of total survey error, combining various sources of error – sampling, non-response and measurement. In so doing, they show that the complexity of expressions for non-response effects increases for more complex parameters (e.g. regression coefficients compared to means), and on the non-response assumptions applied.

When non-response occurs, it is important to consider its impact as one of several sources of survey error that may affect the estimation of population characteristics and the reporting of results. Selecting the best methods for managing non-response at the estimation stage may involve a careful examination of the factors that appeared to influence the non-response during the survey preparation and data collection stages.

The methods for dealing with non-response can be divided into three broad types: 1) methods taken during survey preparation and data collection to maximize participation; 2) additional data collection to address the incidence of non-response; and 3) methods of survey estimation to take account of non-response that has occurred.

2.3 Accessibility and Amenability

Brick and Montaquila (2009) describe two broad classes of non-response, those relating to accessibility and issues with amenability. Accessibility refers to the ability to make contact with the sampled unit. The success of household surveys can depend on finding selected units at the household for example. Similarly, telephone based surveys can struggle to make contact with selected units. When contact is made, issues related to the unit’s willingness to respond are classed as amenability related issues, and a
considerable literature exists exploring various factors that may influence this amenability to respond, for example Beatty and Hermann (2002); Johnson, O’Rourke, Burris and Owens (2002); Willimack, Nichols and Sudman (2002).

In the context of educational surveys, Sturgis et al. (2006) report on outcomes of a survey of teachers from sampled schools that did not participate in the PISA 2003 survey following the removal of the UK data from the international comparison tables of outcomes produced from that survey. They discuss the important issue in the context of institutional surveys (such as surveys of students within schools), of the work to contact and persuade the establishment ‘gatekeeper’:

“Most of the time, initial contact will need to be made with an employee who has both the power to authorise institutional participation and access to the requisite information. Such individuals are usually senior, with little time available for what might be considered ‘non-essential’ activities.”

2.4 Managing non-response during design, preparation and data collection

Considerable efforts will often be made in the survey preparation and data collection stages of the survey to maximize the cooperation of sampled units. Careful consideration needs to be directed towards all aspects of the survey operation - the production of materials, the timing of the survey, the nature of approaches made to sampled units, and the methods used for following up units who have not responded – to ensure that participation in the survey is maximized. Efforts need to be targeted towards addressing likely sources of non-response.
Simply increasing the sample size without improving the representation of those inclined not to respond to the survey may not address the underlying bias. Increasing response rates without consideration of the factors underlying non-response may in fact increase the effects of non-response bias: Groves (2006) notes that “…there can be increases in non-response bias with increasing response rates when persons with distinctive values on the survey variable are differentially sensitive to the design feature creating higher response propensities.” He cites an example of an exit poll where the incentive of a pen was used to increase response rates. While the incentive increased response propensities, it do so differentially between Democrats and Republicans, leading to larger non-response bias in the survey outcomes.

2.4.1 Observations from previous surveys
Lessler and Kalsbeek (1992) outline a series of options and methods available for anticipating the likelihood of non-response in the survey preparation stage, addressing either or both of the two key factors leading to survey bias – the expected response rate and the degree to which respondents and non-respondents are expected to differ on the survey outcome variables. For example, they discuss using identification studies to identify the potential for non-response bias: “the researcher can qualitatively assess the impact of non-response on survey estimates by investigating patterns of response rates among subgroups of the sample”.

An important issue in relation to many educational surveys is that non-response may occur at the two stages of selection that are typically part of the design for these surveys – schools, then students. The drivers of non-response may be different at each stage of
selection. The factors leading to non-response may also be different between schools of different types.

In their study on ‘Which Schools and Pupils Respond to Educational Achievement Surveys’, Schnepf, Durrant, and Micklewright (2014) use logistic and multilevel regression analyses to explore school and pupil response in relation to England’s participation on two PISA surveys. Making use of nationally available achievement and socio-economic background data, they find that average achievement at the school was a less important factor in explaining school non-response than the socio-economic profile of the school. Factors such as location, gender composition, school size and school type did not explain school non-response. In contrast, they found that at the pupil level, student ability was the strongest predictor for explaining non-response, while the socio-economic background factor was not an important factor. They observe that the school and pupil non-response patterns were quite different to each other … ‘indicating that any kind of survey design needs to consider different response mechanisms at both levels for achieving the best possible representative sample’. They note implications with respect to various non-response strategies such as the use of school substitution, discussed later in the thesis (see section 2.6.5). One factor leading to non-response in PISA 2003 that was cited in England’s national report following the 2006 PISA assessment was the timing of the assessment in the school year, in particular problems for schools in the overlap with national exams. For this reason, for future PISA surveys, England Wales and Northern Island were permitted to test outside the normal testing window between March and August. For these jurisdictions the testing took place in November to December (Bradshaw et al (2007)).
Piesse and Rust (2003) discuss response patterns across school types in relation to the United States’ participation in the 2001 Progress in International Reading Literacy Study (PIRLS), finding no statistically significant relationship between response status and the majority of school characteristics that were available amongst the participating schools. However for the original sample of schools, they found that across some characteristics of school authority and funding there were differences in response propensity. They also note some issues relating to differences in school size between original sampled schools and designated substitutes. They conclude that the “use of replacement schools did however seem to introduce a nonresponse bias that was not present in the original sample of schools”.

The findings from these studies illustrate the complexity of non-response for educational surveys, with factors responsible for survey non-response varying at different stages of selection and across different school types. They also illustrate that management approaches (e.g. school substitution) may work better for some parts of the population compared to others. They emphasise the importance of ensuring that good auxiliary information is included at both school and student level in anticipation of non-response for these surveys.

2.4.2 Follow-ups during data collection
A standard practice for many large scale surveys is to follow up selected units that have failed to respond. In relation to student surveys, the operations chapter of the Technical Report for PISA survey (OECD (2012)), describes a process of conducting follow-up sessions so that students absent from the original session complete the assessment.
In the design of follow-up procedures, consideration of the potential for differences between initial responders and those who have not yet responded with respect to the survey outcomes being estimated is required. For example follow-ups with selected units might occur at different times of the day or different days of the week (Brick and Montaquila (2009)) to maximise the response propensity across sub-populations classified by variables (e.g. employment) that might be correlated with survey outcomes.

2.4.3 Two-phased sampling

Beyond operational approaches to follow-up during data collection the survey design may include sample-based approaches towards targeting follow-ups – the selection of a sample of non-respondents that can be used in non-response adjustments or at least to provide estimates of non-response bias.

Observations from previous studies and other substantive knowledge may be drawn upon to inform the design for a future survey. For example a survey design involving two phases of sampling might be planned in the context of the overall survey design to address anticipated non-response. The second phase of data is targeted towards the sampled units that failed to respond to the initial data collection stage. Typically additional resources are employed for this second approach. For example, the first approach might have involved a mail-out questionnaire, but in the second phase of sampling more personalised approaches such as phone calls or site visits might be attempted to try to win the cooperation of the non-respondents sampled for a second
time. In the best case scenario, the additional efforts achieve full response from the group of non-respondents sampled for a second time.

This approach extends beyond the following up of non-responding units that will typically be conducted within the context of a data collection activity. The second stage of data collection might involve the design of separate instruments and alternative operational procedures designed to achieve better success with units who were not disposed to respond to the initial data collection. The second data collection might for example involve a much shorter instrument, so less of the unit’s time is being imposed upon. Or fewer of the more difficult or sensitive items might be included.

A two-phased sampling methodology needs to be planned for from the very beginning. The additional instrument preparation and operational expenses need to be anticipated in the overall survey budget, as well as the extra sample design and estimation work required. Units sampled at the second phase might for example be visited personally compared to the initial stage where instruments were mailed to units, and so the resource requirements per unit may be very different between the two data collection stages. The extent to which the (generally more expensive) second data collection can be resourced will need to be factored in from the very beginning.

One of the features of this approach is the opportunity to collect auxiliary variables cheaply from a large sample of the population at the first phase, which can be used to refine estimates of other variables that are more difficult and expensive to collect and which are therefore restricted to a smaller sample of units taken in the second phase. The use of ‘short’ and ‘long’ forms in the US census is an example of this approach.
Stackhouse and Brady (2003) discuss these forms of the US census and the effects of response rates of having the shorter form. Rao (1995) develops models for determining optimal values of sample sizes and sub-sampling fractions, taking into account the respective costs of collecting data for each phase of the survey.

Bethlehem (2009) describes an alternative basic question approach for situations where a second data collection phase is not possible because of for example time or budget constraints. In this approach the main goals of the survey are condensed into a small number of basic questions. “The basic question approach was born from the observation that people who refuse to participate can often be persuaded to answer a few basic questions”. With this approach, the extent of differences between respondents and non-respondents with respect to the main survey questions can be estimated.

2.4.4 The collection of additional data

Additional data may be collected in order to address anticipated non-response. As already noted, while in the case of unit non-response no information is collected directly from the sampled unit in the course of the survey some information about that unit may still be obtainable. In the case of a sample of students selected in a two-stage process where schools have been sampled at the first stage, information such as the location and socio-economic profile of the school, and the age, year level, gender, socio-economic background and prior performance history of the sampled student may still be obtainable. As will be explored in the course of this thesis, these auxiliary data may play a vital role in statistically adjusting for non-response. They may be key to both analysing the extent of non-response bias in the collected data and to addressing its effects through various adjustment processes in estimation. It is therefore vital that the
anticipation of the likely needs with respect to managing non-response are addressed at the survey preparation stage, so that for example additional items are added to survey instruments specifically for the purpose of addressing anticipated non-response, or more detailed information is collected about the students’ schools than might otherwise have been thought necessary, for example the proportion of students from Aboriginal and Torres Strait Islander (ATSI) backgrounds, the proportion of students with language backgrounds other than English, or the average outcomes on a similar survey conducted in a previous year.

2.4.5 Substitution

Another type of potentially useful additional data to compensate for non-responding units may be obtained via alternative, substitute units. For example, in the case where the students sampled from a school are not obtainable because the school has declined to participate in the survey, it may be possible to use students from a similar school as substitutes for the non-responding students. As discussed at section 2.6.5 school substitution can be considered a type of donor imputation.

The use of replacement schools is discussed in Murphy and Schulz (2006). Durrant and Schnepf (2017) note that “As is common with international surveys on children’s learning achievement, the PISA sample design uses a system of ‘replacement’ of non-responding schools.” School substitution has been a standard practice in Australian educational surveys for many years, and also within international educational surveys in which Australia has participated, e.g. Joncas (2012). For an extended discussion on the use of replacement schools see Prais (2003), Adams (2003) and Sturgis et al. (2006).
2.5 Estimation in the presence of unit non-response

In spite of all efforts at the preparation and data collection stages to minimise both the amount of non-response and its impact with respect to precision and bias, some non-response will likely have occurred that needs further management at the estimation stage following data collection. Management at this stage will include procedures such as weighting to ensure that responding units appropriately represent the parts of the population from which they were sampled, the incorporation of terms into estimators to reflect non-response under certain assumptions, and making use of auxiliary variables and information about non-respondents available for other data sources to model the relationship between survey variables and outcomes. The following sections outline a range of approaches to estimation in the presence of non-response that have developed over recent decades.

2.5.1 The potential impact of non-response bias on survey estimation

Cochran (1953) presents a useful framework for considering the potential size of bias arising from non-response. Assume the population is divided into two strata – the respondents and the non-respondents, and let $W_1 (=N_1/N)$ and $W_2 (=N_2/N)$ represent the proportions of the population in each stratum. Then the population mean for a particular survey variable $\bar{Y}$ will be equal to $W_1 \bar{Y}_1 + W_2 \bar{Y}_2$, that is the weighted sum of the population means for the respondents and non-respondents respectively. When a simple random sample is drawn from this population, the sample mean will be based on the mean of the respondents ($\bar{Y}_1$), and the bias can be written as
\[ E(\bar{Y}) - \bar{Y} = \bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (W_1\bar{Y}_1 + W_2\bar{Y}_2) = W_2\bar{Y}_1 - W_2\bar{Y}_2 = W_2(\bar{Y}_1 - \bar{Y}_2). \]

The bias is composed of the product of two quantities: the proportion of non-respondents \( W_2 \) and the difference in means between the respondents and non-respondents \( (\bar{Y}_1 - \bar{Y}_2) \). It will be small if there is not much difference between respondents and non-respondents with respect to the outcome measures of the survey, and/or if the non-response rate is low. As discussed earlier, considerable efforts will often be made to minimize the rate of non-response during data collection to control the first factor in this expression for the bias. With respect to the second factor, as no data is collected from the non-respondents, it can be very difficult to accurately estimate the differences between the two groups. Estimating the extent of bias for a survey can be a substantial challenge in the estimation phase of the survey. One of the primary reasons for approaches such as two-phased sampling (section 2.4.3) is to obtain an estimate of the non-respondent mean, either directly or via the use of auxiliary information.

Even small rates of non-response can potentially have a large impact on the precision of survey estimates. For example, assume a simple random sample of 1000 people is targeted, and data from 900 of these is collected, representing a response rate of 90%. On one variable, 50% (450) of the sample were found to be in favour of a particular proposition.

If the assumption is made of no difference between respondents and non-respondents, and ignoring any finite population correction, then the standard error of the population
estimate for this statistic would be calculated as $\sqrt{(0.5)(0.5)/900} = 0.017$, and a 95% confidence interval of approximately $0.5 \pm 2 \times 0.017 = (0.47, 0.53)$ would be generated.

However for most surveys, the assumption of no differences between respondents and non-respondents is unrealistic. Many studies have shown that non-respondents tend to differ to respondents with respect to the outcome variables of surveys. Bethlehem (2009) for example describes a follow-up study of non-respondents to the Dutch Labor Force Survey in 2005 finding differences between respondents and non-respondents with respect to location, access to a landline telephone and ethnicity.

In order to set boundaries around the range of possible effects of non-response on inference, Cochran (1953) proposes the following approach:

- When estimating the lower confidence interval limit, assume all non-respondents would have been against the proposition
- When estimating the upper confidence interval limit, assume all non-respondents would have been in favour.

For the lower confidence limit, the sampling proportion assumed becomes 0.45 (450/1000), with a standard error (leaving aside the potential bias in the variance estimator as discussed above) equal to $\sqrt{(0.45)(0.55)/1000} = 0.016$ and the lower level limit generated would be $0.45 - 2 \times 0.016 = 0.42$.

For the upper confidence limit, the sampling proportion becomes 0.55 (550/1000), the standard error is as above, and the upper level limit generated is $0.55 + 2 \times 0.016 = 0.58$. 
The interval (0.42, 0.58) derived from extracting the lower and upper boundaries under these assumptions suggest a potentially substantially less precise outcome than that calculated when an assumption of no difference between respondents and non-respondents is made. To put it another way, it is an interval which would be achieved from a sample size of just 160 if a 100% response rate was obtained

\[ \sqrt{0.5(0.5)/160} = 0.04;0.5 \pm 2 \times 0.04 = (0.42,0.58) \]. Very large differences in outcomes are observed under different assumptions with regard to non-response – the assumption of no difference between respondents and non-respondents, and the assumption of ‘maximum difference’. Confidence intervals calculated using this approach would become very large indeed for high rates of non-response.

The above example simply illustrates the potential impact of non-response on survey estimation. It highlights the necessity for exploring estimation methods that appropriately address the incidence of non-response with reasonable and testable methods and assumptions about the likely effects of missing data on inferences.

2.5.2 Addressing potential bias through a two-phased sampling design

One approach to addressing the potential bias from non-response is to use a two-phased sampling design. The first phase involves a sample size of \( n \) comprising \( n_1 \) respondents and \( n_2 \) non-respondents. In the second phase, a simple random sample of \( m_2 \) of the \( n_2 \) non-respondents is selected. Then an estimator for the population mean is given by

\[
\hat{\mu} = w_1 \bar{y}_1 + w_2 \bar{y}_{2m}
\]
where \( w_1 = \frac{n_1}{n} \), \( w_2 = \frac{n_2}{n} \), and \( \bar{y}_{2m} \) is the sample mean for the second phase sampling of \( m_2 \) of the \( n_2 \) non-respondents. Assuming full response to the second phase sampling, \( \bar{y}_{2m} \) is an unbiased estimator of \( \bar{Y}_2 \), the mean of the non-respondents to the initial sample, and (as \( \bar{y}_1 \) is an unbiased estimator of the mean of the respondents to the initial sample) \( \hat{Y} \) is unbiased.

Rao (1995) shows that the variance of this estimator is given by

\[
V(\hat{Y}) = \frac{(1 - f)}{n} S^2 + W_2 \frac{(k - 1)}{n} S_2^2
\]

where \( f \) is the fraction of the population sampled in the phase 1 sampling, \( S^2 \) is the population variance, \( W_2 \) is the proportion of non-respondents in the population, \( k = \frac{n_2}{m_2} \), the ratio of the number of non-respondents at phase 1 to the size of the sub-sample at phase 2, and \( S_2^2 \) is the variance of the non-respondents. The second term in this expression represents the addition to the variance due to the sub-sampling at the second phase. He derives an estimator for this variance making use of \( s_1^2 \) and \( s_{2m}^2 \), the sample variances of the \( n_1 \) and \( m_2 \) units respectively.

Even with additional resources devoted to approaching the non-respondents compared to the initial approach, in practice achieving 100% response in the second phase is unlikely, and so further adjustments may be necessary, as discussed further in the following sections.
2.5.3 A theoretical framework for non-response management

Assume that from the \( n \) units initially sampled for the survey, responses are obtained from \( n_1 \) units, so that a model needs to be developed to address the missing data from the \( n-n_1 \) non-responding units.

If a survey involves the collection of \( p \) variables from the \( n \) sampled units, then with no non-response, the dataset consists of an \( n \times p \) matrix. Call this matrix \( Y = (y_{ij}) \) and let \( y_{ij} \) be the response to variable \( j \) from unit \( i \). In the case of unit non-response, the rows of this matrix that represent the non-respondents will consist of missing values. Some variables will typically be available even for the non-respondents, for example those that are available from the sampling frame. In the case of student surveys for example, such variables would include the school type and location at the school level, and in some cases information such as the age and gender of the sampled students. These are the sample design and other auxiliary variables, \( Z \), assumed known for all sampled units, and not directly collected from the units themselves.

Little and Rubin (2014) outline a framework - first described in Rubin (1976), and commonly used since - where different types of non-response are categorised by comparing the distribution of the missing data with the distribution of the variables of interest to the survey, as follows:

Define an additional matrix, \( M = (m_{ij}) \) such that \( m_{ij} = 1 \) if \( y_{ij} \) is missing and \( =0 \) if \( y_{ij} \) is present. This is referred to as the missing data indicator matrix. Now consider the distribution of \( M \) conditional on the survey outcomes and on unknown parameters, 0,
external to the survey, \( f(M|Y, \theta) \). If the distribution of missing data does not depend on the outcome variables, \( Y \), or the sample design variables, \( Z \), i.e. if:

\[
f(M|Y, Z, \theta) = f(M|\theta) \text{ for all } Y, Z, \theta
\]

then the missing data are *Missing Completely at Random* (MCAR).

For example, if a student’s non-participation in a test is completely unrelated to ability, gender, or to school or home background factors then the non-participation could be categorised this way. While for most surveys, this is a strong assumption to make about non-participants, it is the assumption made at least implicitly for all surveys where no treatment to address non-response is made. The respondents to the survey are simply assumed to be representative of the population. When data are MCAR, non-participants can be considered as having been selected at random from the sample, and there would be no non-response bias, the respondents to the sample would be representative of the population.

A less restrictive assumption is that the distribution of missing data is related only to design variables, but unrelated to outcome variables, i.e.

\[
f(M|Y, Z, \theta) = f(M|Z, \theta) \text{ for all } Y, \theta
\]

then a model incorporating the design variables can be constructed to account for the non-response. In this case the data are called *missing at random* (MAR). For example, the pattern of non-participation might differ between urban and rural students, but
within each group separately be random. In this case it is straightforward to adjust estimates taking into account the differences for urban students compared to rural students, to achieve unbiased estimates provided we have available a variable that indicates whether the student is urban or rural.

If a relationship between the distribution of missing values and the survey outcome variables exists that cannot be explained away by conditioning on design variables, then the missing data are not missing at random (NMAR). For example, if low ability students are less likely to participate in a survey assessment, and this cannot be fully explained by available auxiliary variables (such as sex, location, socio economic background), then even within subpopulations defined by these conditioning variables, the probability of participation is not uniform between units – i.e. the distribution of unit-missing data is not random – then inferences based on the random selection of units into the sample will be biased.

The selection of the best treatment for managing non-response depends on what can be assumed about the distribution of the missing data. Several approaches assume that the missing data are MAR, so that the non-respondents can be assumed to be a random selection of population units, if not for the population as a whole (where an assumption of MCAR can be upheld), then at least for subpopulations as defined by available auxiliary variables. If the data are found to be NMAR, then further assumptions and/or more complex models will be required to specify the relationship between non-respondents and the survey variables. Methods that are applied under the various assumptions about the missing data mechanism are discussed further below.
2.5.4 Incorporating adjustment terms into estimators

Under full response, a standard estimator for a total of a probability-based sample is the so-called Horvitz-Thompson estimator $\hat{Y}_{HT}$. This is calculated by summing the observed values from the sample weighted by their selection probability:

$$\hat{Y}_{HT} = \sum_{i \in S} d_i y_i$$  \hspace{1cm} 2-6

with $d_i = \frac{1}{\pi_i}$, i.e. the inverse of the selection probability of sample element $i$ (e.g. Maiti (2011)).

Under simple random sampling for example, with equal probability selection of $n$ units into the sample from $N$ units in the population, and assuming full response, $\pi_i = \frac{n}{N}$ and $d_i = \pi_i^{-1} = \frac{N}{n}$.

When non-response has occurred, an adjustment to the design weight might be calculated and added to the estimator. Under the simplest model where non-response is assumed MCAR across the whole population (section 2.6.2), the assumed model characterising the selection of respondents from the sampled units has, for that component, probability $m/n$. The overall selection probability is the product of these two components: $n/N \times m/n = m/N$. As with the design weight, the adjustment weight $(c_i)$ is the inverse of the assumed selection probability under the model. The adjustment to the Horvitz Thompson estimator under this model of non-response might be written:
\[ \hat{Y}_{HTadj} = \sum_{i \in S} d_i c_i y_i = \sum_{i \in S} \frac{N}{n} \frac{n}{m} y_i = \sum_{i \in S} \frac{N}{m} y_i \]

In general, this factor \( (d_i c_i) \) combining the design weight - reflecting the initial selection probability- and an adjustment weight - reflecting the non-response component (and its management)- makes its way into the expressions for expected value and variance. The product \( d_i c_i \) is sometimes denoted by \( w_i \), i.e. \( w_i = d_i c_i \). An estimator weighted by the product of the design weight and a non-response adjustment weight in this fashion is sometimes referred to as a double expansion estimator, e.g. Haziza and Beaumont (2017)

Under broader designs with non-equal selection probabilities these two products of design weight and adjustment will be evident, with \( \pi_i \) and \( \varphi_i \) typically denoted for the respective probabilities.

2.5.5 Response probabilities and their correlation with outcomes

While the approach described by Cochran in section 2.5.1 is useful for evaluating the potential impact of bias arising from non-response, it relied on being able to group sampled units into distinct groups of respondents and non-respondents. In practice this distinction is not typically so simple.

The decision to respond to a survey will frequently depend on a range of survey related factors, such as the timing of the survey, the skills and qualities of the practitioner in
recruiting the unit, or the use of incentives. A younger person may for example be less inclined to respond to a survey where the interviewer is much older. The efforts that survey agencies make to encourage participation are an acknowledgement that the decision to respond will vary for population units. The responses to survey items themselves may also change under differing survey conditions, for example the opportunity for incentives might lead to more favourable responses to survey questions than would otherwise be the case.

The relationship between response rates and bias can be very complex: “…if the characteristics of non-respondents become more distinctive (i.e. an unusual subpopulation on the y variable) as the nonresponse rate becomes small, there may be no reduction in the nonresponse error with lower nonresponse rates” (Groves et al. (2011)).

Bethlehem (2009) gives an expression for the approximate bias under the model described in section 2.5.4 for simple random sampling, which can be expressed as

\[ YC_pC_YR_{pY} \]

where \( C_p \) and \( C_Y \) are the coefficients of variation of the response probabilities and the outcome variable, respectively, and \( R_{pY} \) is the correlation between them in the population. This expression for the bias of \( \hat{Y}_{HT} \) also applies for unequal probability designs (Särndal, Swensson, and Wretman (1992)). This expression clearly shows that the bias is affected
by the variation in the response probabilities and their correlation with the outcome variable.

2.5.6 Design-based versus model-based approaches to inference

The conventional approach to developing the sample design and estimation procedures is based on probability sampling where each unit in the finite population, has a known and non-zero chance of selection. Randomness is solely associated with the variables that indicate whether a population unit is selected in the sample or not. Inference about functions of the population values is based on the randomisation distribution induced by the selection probabilities. Lohr (1999) describes this as a design-based approach.

In the design-based approach estimation is built on the result that weighting by the inverse of the selection probability produces estimates of population totals that are unbiased over repeated sampling from the finite population. Inference rests on the assumption that the sampling distribution of an estimator is approximately normal. Variance estimators can be obtained using explicit formula or replication methods such as the Jackknife (Wolter (2007)).

In the model-assisted approach (Särndal et al., 1992) statistical models for the population values are used to motivate sample design and estimation but the inferences are still based on the randomisation distribution. In the model-based approach a statistical model for the population values is used explicitly in design and estimation (Chambers and Clark, 2012). Statistical properties of estimators are derived with respect to the assumed model
conditional on the selected sample. The issue of how to select the sample still needs to be considered. Probability sampling is not required but it may be used. Use of a non-informative sampling scheme is usually assumed, where the sample selection depends on variables that are available before the sample is selected and there is then no bias conditional on these variables.

Rubin (1983) for example describes a Bayesian approach where a posterior probability distribution for the unknown population values using the observed values is estimated, and this distribution is used to make inferences. So long as the model is well specified, model-based estimation can lead to estimates with lower variance than those obtained under a design-based approach. If the model explaining the observations can be well specified this approach may be particularly appropriate for analysing data known to be particularly skewed, for which a design-based framework may result in unnecessarily large variances around survey estimates. However, the approach is very sensitive to model misspecification. Rubin (1983) notes “…the Bayesian must always be concerned about the possibility of an inappropriate answer if the specified model is not a good reflection of reality. Randomisation inference, on the other hand, avoids a particular model specification for Pr(Y) and consequently tends to be suboptimal but more robust.”

These approaches are frequentist in that they consider properties of estimators over the repeated sampling or frequency basis (Welsh, 2011). They differ in that the design-based approach considers the repeated sampling from a fixed population of values whereas the model-based approach considers repeated realisation of the values of the selected units from a probability distribution.
2.6 Non-response adjustment approaches at the estimation stage

The following sections describe a range of approaches for the management of non-response at the estimation stage frequently observed in practice. They being with weighting adjustments made under simple assumptions about the mechanisms underlying non-response, either for the population as a whole or for identifiable sub-groups of the population. Following these are approaches that make use of auxiliary information collected from the survey or available from outside data sources. These include using these data to estimate the response propensity, or incorporating these data into estimators or models that take account of the relationships between collected data, outcomes and response.

2.6.1 Weight classes and post-stratification

Weighting class adjustments and post-stratification are very widely used approaches to managing non-response. In post-stratification, the respondent data set is weighted to represent population subgroups where the sub-population sizes are known. For weighting class adjustments, the sub-population subgroups are estimated rather than known. Brick and Montaquila (2009) and Haziza and Beaumont (2017) present reviews of approaches to the construction of weights that are used in practice, including basic weighting systems applied to all estimation for a survey and more tailored approaches where, for example, certain available auxiliary variables are applied to particular estimates and other auxiliary variables for other estimates. They also discuss approaches to the trimming and smoothing of weights, and the application of weighting to domain estimation.
2.6.2 Quasi-randomisation

A commonly adopted assumption under post-stratification or weighting class adjustment is that all population units have a positive probability of being a respondent, and that subgroups of the population can be constructed (weighting classes) within which the probability of response can be assumed to be equal for all units. The approach is equivalent to considering the decision to respond to the survey to be a final stage in the sample selection process. Oh and Scheuren (1983) use the label ‘quasi-randomisation’ to describe this approach, with the ‘quasi’ referring to the implicit modelling of the response mechanism that underlies the approach, and ‘randomisation’ referring to the fact that the estimators and inference developed in this approach are based on the sample selection probabilities of the design – the traditional methods for drawing inferences from probability-based samples.

Inevitably the treatments required for achieving satisfactory variance estimates need to be tailored to the particular survey taking factors such as the differences in weighting class means, the weighting class sample size, and the respondent size within each weighting class into account.

Oh and Scheuren (1983) note some dangers associated with the uncritical use of the derived expressions. One danger is that the respondent sample size for some weighting classes within which the assumption of a uniform response mechanism can be assumed may become quite small, leading to potentially very large weight adjustments and
inflated variance estimates. The second is that where some weighting classes contribute most of the variance, and the number of responding units from these classes is small, the assumption of normality of the estimator might not be safe.

Another obvious difficulty is that there will generally be more than one variable of interest for a survey. Weighting classes or post-stratification adjustments that appear to work well for one outcome may not be optimal for other outcomes being estimated. It would be possible to construct different weighting classes for different variables, however the practicalities are that a single set of weighting classes will usually be formed for estimation across a range of survey outcomes, and so across variables, compromises are likely to be required.

In practice weighting class and post-stratified estimators involve judgements which balance the desire on the one hand to reduce bias by forming classes within which a uniform response mechanism can be safely assumed, with the problem of small weighting classes leading to larger, more unstable variance estimates. In some cases a small number of units in the weighting class may be given very large weights relative to other units in the sub-population with the risk that they may unduly influence estimates of outcomes. The approach of trimming weights where there is significant variation in the weights is discussed further in section 2.6.13.

2.6.3 Response homogeneity groups
An extension of the quasi-randomization approach is to use the responding sample to form subpopulations of the population – response homogeneity groups (RHGs) - within which a uniform probability of response can be assumed. Unlike the quasi-randomisation model, the subpopulations are not assumed fixed, but can vary with each realised sample. A different number of groups might be formed for one realized sample versus another. Särndal et al. (1992) note with respect to RHGs: “No practitioner really believes that all elements in a group have exactly the same probability to respond, but the point is that the assumption of constant probability within well-constructed groups removes most of the non-response bias”

As with Oh and Scheuren’s (1983) quasi-randomisation models, the framework for response homogeneity groups naturally extends beyond models where the data is considered MCAR within defined RHG’s, to models where the data are considered MAR, conditional on the design variables that form the RHG.

Särndal et al. (1992) demonstrates that within the response homogeneity framework, estimators that make use of auxiliary information have, in addition to the well-known benefit of reduced variance derived from the correlation between auxiliary information and the outcome variables of interest, the additional (and more important) benefit of providing improved resistance against non-response bias when the assumed model (e.g. MAR within RHGs) is erroneous.

2.6.4 Extended probability sampling approach

The approaches discussed above involve some sort of model to estimate the response mechanism, e.g. sub-populations are defined within which the propensity for units to
respond to the survey is assumed to be equal. These approaches involve two sources of randomness – the randomness associated with the selection of units into the sample according to the sample design, and the randomness associated with the propensity to respond – the ‘response mechanism’. The population is assumed to be fixed and finite, and inferences are made with reference to the randomisation properties of the sampling distributions of estimates.

More broadly, Särndal et al. (1992) refers to an ‘extended probability sampling approach’, involving (a) the randomisation distribution induced by probability sampling; and (b) one or more model assumptions to address non-sampling errors, including (but not limited to) those arising from non-response. They write: “The known randomisation distribution continues to play an important part in the inferences. But … the conclusions about the finite population depend, for their validity, on the truth of [the models invoked to address the non-sampling errors]. We no longer have the distribution-free property that pure probability sampling theory prides itself on”.

2.6.5 Imputation

Imputation is another approach for managing non-response at the estimation stage. It involves substituting a synthetic value for the missing value. The substitute value may still be a real value from another hopefully similar unit in the data set selected in some way or may be generated using some imputation model or methods that exploit the possible relationship between the unobserved items and the observed ones. There are a number of imputation methods applied in sample surveys as reviewed by Haziza (2009) and Lohr (1999). Imputation methods are frequently used for the management of item
level non-response. Although not common, imputation methods can also be used to handle unit nonresponse, including with the use of donor imputation approaches and in the application of imputation models, as described below.

**Donor imputation**

A range of donor imputation methods have been developed to address unit level non-response. In these approaches, a separate record may be substituted for the non-responding unit.

One approach is so-called ‘hot-deck’ imputation where the value (or set of values) from one of the responding units is substituted for the missing values. This can be done sequentially, randomly, or a ‘near neighbour’ approach, where a measure of distance of records to the unit in question is calculated, for example on the basis of known covariates, or other substantive knowledge, and the record with closest distance is used in the imputation. Also there can be ‘cold-deck’ imputations where imputed values are derived from historical data.

**School substitution**

A special case of ‘near neighbour’ donor imputation used in many educational surveys, including those conducted under Australia’s National Assessment Program (NAP), has been the practice to use substitute schools for sampled schools that choose not to participate. At the time of sample selection, one or two substitute schools are identified that are similar to the sampled school with respect to stratification and size variables.
that have been used in the sampling process that are known to be correlated with the major outcome variables of the survey. If the sampled school declines the request to participate in the survey, then one of its designated substitutes may participate in its place. The data collected from the substitute is used to represent the data that would have been collected from the sampled school.

The underlying assumption with the use of substitutions is that the students from the substitute school can be used as unbiased representatives of the students from the non-responding school. In one respect at least, i.e. the preparedness to respond to the survey, an included replacement school has demonstrated its difference from the sampled school that it is replacing. Some practitioners use this as a basis to argue that while substitution will improve the yield of the sample, the use of substitutes does little to remove non-response bias. However, the same argument applies to weighting and other imputation methods described above. Both approaches attempt to account for the data not collected from the non-respondents by using data from other population units. Chapman (1983) writes that “the key question regarding the worth of substitution procedures is whether the use of substitutes provides better proxy values for non-respondents than those provided by alternative imputation procedures”.

The relative merits of substitution depend on the amount of information available about the non-respondents, and also the size of the sample within the part of the population where the non-response has occurred. Chapman (1983) suggests that in surveys of institutions (including schools), where a substantial amount of stratification information is available about the non-responding schools, but where the sample sizes within subclasses of the population are necessarily smaller because of the relatively high costs
associated with surveying through institutions, the use of substitutions may produce better imputations than other weight adjustment procedures.

It is possible to apply substitutions to different levels of sampling. Most student surveys involve two stages of sampling, first the selection of schools and second the selection of students within the school. As discussed above, substitutions might be used to replace non-responding schools. It would also be possible to substitute non-responding students within the sampled school with other students not originally sampled. A problem with substitution at this stage of sampling is that it is more difficult to identify a ‘like student’ to replace the non-responding student. At best, information such as gender and year level might be available from which a similar student with respect to these variables might be selected. However, across the schools surveyed to participate, it is difficult to know whether replacing a non-responding student with a student in this way will achieve a better outcome than using the responding students who were sampled to represent all students who were sampled from the school. Substitutions of students also add more burden to the field work. For example, parental permission may need to be collected in advance from potential substitutes to account for student non-response that presents at a late stage. Substitute students may need to be located at short notice; participation coding needs to be applied that clearly identifies that the student was a substitute and so on. For these reasons, most student surveys have not used substitution of students in their data collection methodology.

There is some risk that the use of substitutions for non-responding units might exacerbate bias in survey estimates. When presented with a reluctant unit, field workers may be tempted to move too readily to the use of a substitute rather than work to win
the cooperation of the sampled unit. Alternatively, the identification of the need for substitutes might occur quite late in the data collection phase and as a result, the procedures for approaching and winning the cooperation of a suitable substitute might differ from the procedures applied to the originally sampled units. Where a shortfall in a response rate standard is encountered, special measures such as incentives, extra follow-up attempts, or solicitations from authorities urging cooperation might be employed to win the cooperation of substitute units within a compressed timeframe, which might lead to different response patterns and therefore potentially increase bias. For these reasons the use of substitutes must be used only in a limited way, and should be quantified as part of the reporting. It is important for field procedures with regard to units to participate, whether originally sampled or substitutes, be as consistent as possible.

**Empirical studies on the effects of substitution**

Chapman (1983) reports on the outcomes of some empirical studies on the effects of substitution on reducing non-response bias in surveys. One such investigation was undertaken for a longitudinal survey conducted in the early 1970s exploring the destinations and attitudes of high school graduates, relating these to the students’ personal and educational background. A two-stage sampling design was used to select 1200 schools, two from each of 600 strata, with 18 students selected from each participating school. At the time of sampling, two additional schools were selected from each stratum to be used as substitutes if required. If neither the sampled school nor one of its designated substitutes agreed to participate the school was treated as a non-respondent. A total of 1649 students from 974 schools participated in the survey,
including 53 schools that were used as substitutes for non-cooperating schools. At the first follow-up stage, as well as surveying once again those that had taken part in the original survey, the researchers undertook a complete follow-up of the schools that did not respond at the first stage.

For the original survey, about 20% of original sampled schools did not participate, but the non-response was reduced to 2% of original sampled schools after the first follow-up. The researchers were then able to compare estimates of totals and proportions for 35 multiple choice items between their estimates from the original survey based on the use of the substitute schools and also weighting adjustments to reflect the remaining non-response with estimates derived after the follow-up study was conducted. They tested each of the 155 response categories from the 35 items for bias, and rejected the null hypothesis of no bias 91 times out of 155. They derived an average estimate of school non-response bias of approximately negative 5%.

The other empirical investigations of bias reported by Chapman (1983) tended to produce the same conclusion, that the substitution procedures did not eliminate the effects of the non-response bias. However, as he points out, even though the use of substitution was not successful in eliminating the non-response bias, other methods might have fared even worse. He recommends that to evaluate the usefulness of substitution procedures, efforts should be continued to obtain the cooperation of the non-respondents, and following this, the use of substitutes can be compared with other methods of adjusting the data to account for non-response.
2.6.6 Imputation models

Because donor imputation is based on an actual record, it can be considered more ‘genuine’ than some outcome derived from a statistical model. However, such methods do not fully exploit the auxiliary data that may be available.

In order to better exploit available auxiliary information, imputation models can also be developed and used. If there are some auxiliary variables available from the sample frame or through some other source for non-responding and responding units then these can be used in the imputation methods or models.

Rather than modelling the response probability and applying weighting adjustments to the respondent data, as described in the above methods, the outcome measure, which is missing due to non-response, is predicted based on available auxiliary data. The quality of the imputation depends on the strength of the relationship of the outcome measures to the auxiliary data.

*Regression imputation*

An example is with the use of regression imputation, where the missing values for outcome variables are imputed using regression models to make a prediction for the outcome based on the observed relationship between auxiliary variables and the outcomes from the sample data. Bethlehem (2009) expresses a class of imputation methods using the following expression
\[
\hat{y}_i = B_0 + \sum_{k=1}^{p} B_k X_{ki} + E_i
\]

\(X_{ki}\) denotes the value of the auxiliary variable \(X_k\) for element \(i\). \(B_0, B_1, \ldots B_p\) are regression coefficients and \(E_i\) is a random term which is generated in varying ways depending on the imputation method. This includes mean imputation, which can be applied over the whole sample, or more commonly within groups defined by some auxiliary variables.

The inclusion of the random error term in the above expression reflects a challenge with imputation which is that the approach needs to somehow capture the uncertainty associated with the prediction in order to properly represent the variation in estimates associated with the method used. Rubin (1986) writes:

...analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reasons for nonresponse are known. Equally serious, single imputation cannot represent any additional uncertainty that arises when the reasons for nonresponse are not known.

**Multiple imputation**

In order to address this limitation in simple regression imputation, and also with advances in statistical computing, it has become very popular in recent years to employ a multiple imputation approach, see for example Rubin (1986) or Berglund (2010). The general approach under multiple imputation is to generate multiple, (say \(m\)) values for each missing data item, (in so doing creating \(m\) complete datasets), to analyse each of
those datasets using standard methods, and then to combine the results of those outcomes in order to generate estimates and variance estimates for inference.

The mean of the $m$ estimates derived from each separate analysis is taken as the combined point estimate for the outcome measure being imputed. The within imputation variance is the average variance within the imputed datasets. The between imputation variance is the variance across the $m$ imputed datasets, and the total variance is the sum of the within and between variances (e.g. Berglund (2010)).

Multiple imputation is particularly useful for coping with item non-response, but may also be used to account for unit non-response, using available auxiliary variables.

2.6.7 Weighting approaches versus imputation
As illustrated above, weighting and imputation methods are two commonly used methods for addressing unit non-response. Weighting methods tend to be used where the main interest is in simple descriptive statistics such as means and totals and the main concern is unit non-response. Imputation can be used for unit non-response or item non-response, although it is more common for the latter. Haziza (2009) notes “although imputation is sometimes used to handle unit non-response, it is mostly used to compensate for item nonresponse.”

2.6.8 Response propensity estimation
An extension of the approach of generating weighting classes or response homogeneity groups within which the propensity to respond is assumed to be equal, is to directly model the individual response propensity.
Cassel, Särndal and Wretman (1983) suggest this as a possible bias reduction technique, particularly in cases where the key auxiliary variables are continuous. They suggest the possibility of a logistic regression with response to the survey as the outcome, regressed on auxiliary variables with values known for all sampled cases to estimate the individual response probability.

Under a logistic regression model for example, the response propensity, \( \varphi_i \), can be estimated using:

\[
\log \left( \frac{\varphi_i(x^*)}{1 - \varphi_i(x^*)} \right) = \alpha + \beta' x^*
\]

The inverse of the estimated response propensities might then be added to the inverse selection probabilities as weights to the adjusted Horvitz Thomson estimator (section 2.5.4) as an estimate for the outcome variable under non-response. A fuller description of the formulation of response propensity estimation is provided in section 2.7.3.

\[
\hat{Y}_{HTprop} = \sum_{i \in S} \pi_i^{-1} \varphi_i^{-1} y_i
\]

While modelling individual response propensities can address the issue of response bias, a limitation with this approach can be larger variances associated with estimates of very low response probabilities (and hence very large weights) for some sampled cases. Little and Rubin (2014) suggest a practical procedure of forming weighting adjustment
classes based on a ‘coarsened’ estimate of the response propensity (into a small number of values) in order to minimise the excessive variances that can occur with estimates based on individual response propensities.

2.6.9 Estimators directly incorporating auxiliary information – generalised regression estimators

The methods described above make use of auxiliary information through the approaches such as the formation of weighting classes or response homogeneity groups within which the probability of response is assumed equal, or through the estimation of propensity via logistic regression or other such models. Another approach is to use estimators that make use of the auxiliary information directly within the estimators of outcomes.

An important class of such estimators, widely used in practice, are ‘generalised regression estimators’. Auxiliary variables are used in a regression model to estimate the vector of regression coefficients for a best fit of the dependent variable, $Y$, on the explanatory variables, $X$. Under simple random sampling without replacement and under full response, it can be shown that the vector of regression coefficients estimated from the model is an asymptotically design unbiased estimator of the coefficient vector under the model using the full population. That is, the bias vanishes for large samples (Bethlehem (2009)). The variance of the generalised regression estimator can be approximated by:

$$V(\bar{y}_{GR}) = \frac{1 - f}{n} s^2_e$$
where $S^2_E$ is the population variance of the residuals (Bethlehem (2009)). This expression is the same as the variance under simple random sampling except that $S^2$, the variance of the population values is replaced by $S^2_E$, the variances of the residuals under the regression model. Under a model where the auxiliary information well explains the behaviour of the target variable, the variance of the residuals will be lower than that of the population values themselves, leading to more precise estimates.

The combination of protection against bias and improved precision as described above mean that generalised regression estimators are widely used for managing non-response.

2.6.10 Multiplicative weighting

Another commonly used approach for incorporating auxiliary data into estimators is the use of multiplicative weighting approaches, including post-stratification estimators and so-called ‘raking’ estimators.

In post-stratification a set of auxiliary qualitative variables across any number of dimensions, for example age-group, sex and location, are cross-tabulated and an iterative procedure is used to generate weights which reproduce the population counts over the set of cross classes formed. As with the response propensity adjustment described in section 2.6.8, the calculated weights are adjustments added to the design weights to produce an adjusted Horvitz Thompson estimator (section 2.5.4).

‘Raking’ is a widely used alternative to post-stratification which can be used when the population counts of each of the cross-classes in the cross tabulation of auxiliary
variables is not known, but the marginal totals of the auxiliary classes are available. For example population counts by age-group and sex are available but not the counts by sex within age-group. An iterative procedure described in Bethlehem (2009) describes the procedure that iteratively adjusts the weight factors until the point where they align across all of the dimensions of the cross classification.

2.6.11 Weight calibration

Drawing on the methods described above to align weights to correlated auxiliary variables, Deville and Särndal (1992) developed a generalised framework for using auxiliary variables to improve the efficiency of estimators, known as ‘calibration weighting’.

The desirable properties under calibration estimation are that the adjustment weights should be as small as possible and that for the vector of auxiliary variables used in the calibration, the weighted sample distribution should match to the population distribution. “The first condition sees to it that resulting estimators are unbiased, or almost unbiased, and the second condition guarantees that the weighted sample is representative with respect to the auxiliary variables used” (Bethlehem (2009)). Methods such as the use of generalised regression, post-stratification or raking based estimators can all be shown to fit under this framework, at least for situations where there is full response. Under non-response, Bethlehem (2009) notes that these methods then depend on the underlying models hold with the auxiliary variables as explanatory variables.
Selecting calibration variables

Whether model-assisted or model-based methods incorporating auxiliary information are adopted, the final choice of auxiliary information to be incorporated into estimation will occur following data collection. As noted by Silva and Skinner (1997), the addition of too many calibration variables can lead to over-specification of the response model and substantial increases to the variance. They suggest an approach for the stepwise selection of calibration variables using sample data. Chambers and Clark (2012) propose an adaptive approach for variable selection within a model-based framework.

Weight trimming

An undesirable effect that can occur under some of the non-response management methods described above is for a large variation in weights to be generated, for example in the estimation of response propensities discussed in section 2.6.8 and in the use of raking methods (section 2.6.10). This can cause instability and increased variances of estimates and the issues associated with weight variation are discussed in section 14.4 of Valliant et al. (2013). The design effect due to weighting, also called the unequal weighting effect, is discussed in Kish (1965) and is given by 1+CV^2, where CV is the coefficient of variation (SD divided by the mean) of the weights. In practice weights can be trimmed when the variation may have an appreciable detrimental effect on estimate. Potter (1990, 1993) develops some weight trimming methods and Valliant et al. (2013, pp388-390), review weight trimming and methods that set upper and lower limits. They give an example of trimming weights that are more than 3.5 times the median weight. Valliant and Dever (2018) discuss weight checking and suggest that action might need be taken when the unequal weighting effect exceeds 3 and identify outliers that are candidates for trimming when they exceed the median weight plus 3 time the
interquartile range or the mean weight plus three times the standard deviation of the weights. In the PISA survey, student level weights that exceed four times the median weight for the stratum are trimmed to that value OECD (2012). If there is a large variation weights and an appreciable proportion of weights are flagged as potential outliers that may suggest that the weighting method should be modified.

Inevitably the treatments required for achieving satisfactory variance estimates need to be tailored to the particular survey taking factors such as the differences in weighting class means, the weighting class sample size, and the respondent size within each weighting class into account.

2.7 Formulation of key methods used in non-response

The review of the literature of the first part of this chapter identified a number of key approaches to the management of non-response that will be applied and compared under different non-response scenarios in the subsequent chapters of this thesis.

The following sections outline the basic formulation of several of the methods that will be explored in this thesis: post-stratification and weighting under quasi-randomisation assumptions, the use of weights reflecting group and individual selection probabilities, the modelling of response propensities and generalised regression estimation. They illustrate the additional considerations that are made with respect to the estimation of precision and bias under non-response conditions.
2.7.1 Additional variance under quasi-randomisation

In the simplest case of a quasi-randomisation approach to non-response management (section 2.6.2), it is assumed that the response probability is equal, positive and independent for all population units. This is equivalent to assuming a MCAR mechanism for response to the survey.

Assume that the survey wishes to estimate the population total \( Y = \sum_{i=1}^{N} y_i \) for some characteristic of interest:

Let D be an indicator of selection of units into the sample:

\[
D_i = \begin{cases} 
1 & \text{ith unit is selected} \\
0 & \text{ith unit is not selected} 
\end{cases}
\]

Under simple random sampling of \( n \) units from the N population units:

\[
\Pr(D_i = 1) = \frac{n}{N} \tag{2-13}
\]

From the sample of \( n \) units, assume that \( m \) respond to the survey. Let R be an indicator of survey response:

\[
R_i = \begin{cases} 
1 & \text{ith unit responds if sampled} \\
0 & \text{ith unit does not respond if sampled} 
\end{cases}
\]

Under the assumption that the response probability is positive and equal for all units:

\[
\Pr(R_i = 1|D_i = 1) = \frac{m}{n} \tag{2-14}
\]
An estimator for the population total $Y$ is

$$\bar{Y} = \frac{N}{m} \sum_{i=1}^{N} D_i R_i Y_i$$  \hspace{3cm} (2-15)$$

\ldots which is unbiased

$$E(\bar{Y}|n,m) = \left( \frac{N}{m} \right) \sum_{i=1}^{N} \left( \frac{n}{N} \right) \left( \frac{m}{n} \right) Y_i = \sum_{i=1}^{N} Y_i = NY$$  \hspace{3cm} (2-16)$$

The variance of $\bar{Y}$, given $m\geq1$, is

$$Var(\bar{Y}|n,m) = N^2 \left( \frac{1}{n} - \frac{1}{N} \right) V + N^2 \left( \frac{1}{m} - \frac{1}{n} \right) V = N^2 \left( \frac{1}{m} - \frac{1}{N} \right) V$$  \hspace{3cm} (2-17)$$

with $V$ equal to the element variance of the population units, i.e.

$$(N - 1)V = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$  \hspace{3cm} (2-18)$$

An unbiased estimator for $V$ under the model is
\[(m - 1)\hat{V} = \sum_{i=1}^{N} D_i R_i Y_i^2 - \frac{1}{m} \left( \sum_{i=1}^{N} D_i R_i Y_i \right)^2 \]

The first component of equation 2-17, \(N^2 \left( \frac{1}{n} - \frac{1}{N} \right) V\), is the standard result for the variance of the estimator assuming full response, while the second component of this expression of the variance, \(N^2 \left( \frac{1}{m} - \frac{1}{n} \right) V\) “is directly attributable to the additional level of sampling introduced by the response mechanism” Oh and Scheuren (1983). Clearly the greater the difference between the respondent sample size, \(m\) and the overall sample size \(n\), the larger the increase in the variance attributable to the non-response. In other words, the variance increases with increasing levels of non-response.

### 2.7.2 Weighting classes and post-stratification

Oh and Scheuren (1983) extend the framework for cases where response probabilities – whilst still independent and positive – are equal within subpopulations but may vary between subpopulations, i.e. the missing data are assumed MAR providing a variable indicating subpopulation membership is available.

For example, if the population is divided into \(H\) subpopulation groups (strata), then making use of the assumed known subpopulation size \((N_h)\), the sample size within each stratum \((n_h)\), and the assumed known number of responses \((m_h)\) in each subpopulation group, then within strata under simple random sampling:

\[\Pr(D_{hi} = 1) = \frac{n_h}{N_h}\]
\[ \Pr(R_{hi} = 1 | D_{hi}) = \frac{m_h}{n_h} \]  

The post stratified estimator for the population total takes the form

\[ \tilde{Y}_{ps} = \sum_1^H \frac{N_h}{m_h} \sum_1^{N_h} D_{hi} R_{hi} Y_{hi} = \sum_1^H \frac{N_h}{m_h} \tilde{Y}_h \]

which is unbiased

\[ E(\tilde{Y}_h | n, m) = \sum_1^H N_h \tilde{Y}_h = Y \]

The conditional variance of this estimator is

\[ V(\tilde{Y}_{ps} | n, m) = \sum_1^H (N_h)^2 \left( 1 - \frac{m_h}{N_h} \right) \frac{V_h}{m_h} \]

…with estimates for \( V_h \) given by

\[ (m_h - 1) V_h = \sum_1^{N_h} D_{hi} R_{hi} Y_{hi}^2 - \frac{1}{m_h} \left( \sum_1^{N_h} D_{hi} R_{hi} Y_{hi} \right)^2 \]
A parallel series of expressions are developed in the case of a weighting class estimator, where the $N_h$ are estimated rather than known. Conditional and unconditional weighting class estimators are developed under these conditions.

In short, the expressions demonstrate that estimation under quasi-randomisation parallel the standard expressions found in the design-based framework, but with an additional component reflecting the model of equal probability subsampling of respondents from the sampled units within strata.

For example, adapting from what Särndal et al. (1992) refer to as a naïve response model where:

$$\Pr(R_i = 1 | D_i = 1) = \varphi_i = \varphi$$  \hspace{1cm} (2-26)

(in other words, MCAR). An estimator of the population total under this model would be:

$$\tilde{Y} = N\tilde{Y} = N \frac{\sum_{i=1}^{N} D_i R_i Y_i (\pi_i \varphi)^{-1}}{\sum_{i=1}^{N} (\pi_i \varphi)^{-1}} = N \frac{\sum_{i=1}^{N} D_i R_i Y_i (\pi_i)^{-1}}{\sum_{i=1}^{N} (\pi_i)^{-1}} \hspace{1cm} (2-27)$$

…which is an analogue of equation 2-15, but allowing for varying selection probabilities.

Särndal et al. (1992) say of this model: “Since the unknown $\varphi$ conveniently vanishes, this estimator can always be calculated. However it corresponds to ‘doing nothing about
the non-response’ in the sense that the response model implies no difference in the weighting of respondent values”. The estimator is described as ‘essentially unbiased’: “The negligible bias is not due to the nonresponse per se, but to the fact that the estimator is of the ratio type”.

2.7.3 Modelling of individual response propensities

Section 2.6.8 describes the estimation of individual response propensities as a method for addressing non-response.

Following sample selection (with selection probability \( \pi_i \) for unit \( i \), and the distinguishing the response propensity of an individual sampled unit as \( \varphi_i \):

\[
\Pr(R_i = 1|D_i = 1) = \varphi_i
\]

the Horvitz-Thompson estimator can be expressed

\[
\hat{\theta}_{HT\text{prop}} = \sum_{i \in S} \pi_i^{-1} \varphi_i^{-1} y_i = \sum_{i \in S} d_i c_i y_i
\]

(with \( \pi_i \) and \( \varphi_i \) assumed positive for all \( i \)).

The response propensity is the probability that the sampled unit will respond to the survey, conditional on the vector of auxiliary variables, \( x^* \) assumed for the model.

\[
\varphi_i(x^*) = \Pr(R_i=1|x^*)
\]
A frequently used logistic model \textit{R. Valliant et al. (2013)} is to assume that the response propensity follows a logistic distribution, with

\[
\log \left( \frac{\phi_i(x^*)}{1 - \phi_i(x^*)} \right) = \alpha + \beta'x^*
\]

Other models, such as probit and complementary log-log models are described in \textit{R. Valliant et al. (2013)}.

2.7.4 The Generalised Regression Estimator

One class of estimators that make use of correlated auxiliary data to improve estimation are known under the collective term, ‘Generalized Regression Estimator’, see for example \textit{Lohr (1999)} . These are an extension of the estimate derived from the Horvitz-Thompson estimator \( \hat{Y}_{HT} \). Taking a vector of values for auxiliary variables \( x_i^* \) assumed known for all elements of the population, and assuming no non-response, the generalized regression (GREG) estimator takes the form:

\[
\hat{Y}_{GREG} = \hat{Y}_{HT} + \left( \sum_{i \in U} x_i^* - \sum_{i \in S} d_i x_i^* \right) \tilde{B}_{s:d}
\]

where

\[
\tilde{B}_{s:d} = \left( \sum_{i \in S} d_i \tau_i x_i^* (x_i^*)' \right)^{-1} \left( \sum_{i \in S} d_i \tau_i x_i^* y_i \right)
\]
is a vector of regression coefficients obtained through the use of regression of $y$ on the values of the vector of auxiliary variable for the elements in the sample. The term $\tau_i$ is inclusive of any adjustments made to the sample design weights $d_i$ in the estimation process. In the standard case the sample design weights are used, $\tau_i = 1$.

$\hat{B}_{s,d}$ can be interpreted as an estimate of the regression coefficients ($\beta$) in the model

$$y_i = x_i'\beta + \epsilon_i$$

with residuals, $\epsilon_i$ with mean 0 and variance $\nu_i$.

In the case of simple random samples without replacement and $\tau_i = 1$, the estimator $\hat{B}$ reduces to

$$\hat{B}_s = \left(\sum_{i \in S} x_i^* (x_i^*)'\right)^{-1} \left(\sum_{i \in S} x_i^* y_i\right)$$

which is the estimator of regression coefficients used in standard linear regression.

(Särndal and Lundström (2005), R. Valliant et al. (2013))

In the case when the generalised regression estimator includes an intercept term and $c_{i=1}$ the estimator can be written as $\left(\sum_{i \in d} x_i^*\right) \hat{B}_{s,d}$ (Särndal et al. (1992), page 234)
Availability of auxiliary data

Estimators within the ‘generalised regression estimators’ class can of course vary according to the auxiliary variables that are included in the estimator. Another important variation is with respect to the availability of the auxiliary data. Expression 2-24 above assumes that the auxiliary information is available for all units in the population. Related estimators can be used when the auxiliary data is known for all sampled units but not all units in the population. In other cases, auxiliary data may be available for all population units on some variables, but for sampled units only on other variables. Särndal and Lundström (2005) describe the possibility of a stacked auxiliary vector \((x_k^*, x_k^0)'\) made up of a combination of variables \(x_k^*\) with information known for all population units, and variables \(x_k^0\) with information known for sampled units only.

Whichever combination of auxiliary variable inputs available, these can be used in the estimation of the vector of regression coefficients within the regression estimator.

The Generalised Regression Estimator when there is non-response.

Under full response, the term in brackets in Equation 2-32: \(\sum_{i\in U} x_i^* - \sum_{i\in S} d_i x_i^*\) is close to a vector of zeros, as the HT estimator of the population total based on the sample data \(\sum_{i\in S} d_i x_i^*\) should be close to the population total for each of the auxiliary variables.

When there has been non-response, the expression \(\sum_{i\in r} d_i x_i^*\), summed over respondents, will be an underestimate of \(\sum_{i\in U} x_i^*\). The design weight, \(d_i\), can be changed by a new factor \(v_i\), such that \(\sum_{i\in r} d_i v_i x_i^*\) becomes a good estimate of the population total \(\sum_{i\in U} x_i^*\).
The raising factor, which depends linearly on \( x_i \) the vector of known values of the auxiliary variables for each student, is given by

\[
v_i = 1 + \lambda' x_i
\]

where

\[
\lambda'_r = \left( X - \sum_{i \in r} d_i x_i \right)' \left( \sum_{i \in r} d_i x_i x_i' \right)^{-1}
\]

2.7.5 Calibration

The outcomes of generalised regression estimation can also be expressed in a manner similar to equation 2-7, as an adjustment weight applied to the design weight in the estimator.

\[
\hat{Y}_{\text{GREG}} = \sum_{i \in s} d_i g_i y_i
\]

with \( d_i = \frac{1}{n_i} \) reflecting the selection probability of unit \( I \), and

\[
g_i = 1 + \lambda'_s c_i x_i^*
\]

the so-called ‘g-weight’ derived through the generalised regression estimation. The column vector \( \lambda'_s \) is

\[
\lambda'_s = \left( \sum_{i \in U} x_i^* - \sum_{i \in s} d_i x_i^* \right)' \left( \sum_{i \in s} d_i c_i x_i^* (x_i^*)' \right)^{-1}
\]
A property of the g-weight is that the sum of the auxiliary variables weighted by the design weights adjusted with the g-weight add to the population totals of those auxiliary variables.

\[ \sum_{i \in s} d_i g_i x_i^* = \sum_{i \in U} x_i^* \] 2-41

The weight product \( d_i g_i \) is not dependent on any particular survey outcome, and can therefore be applied across a range of survey variables to generate population estimates – a useful property in surveys where many outcomes are estimated and are said to have been “calibrated to the input of information, the population total \( \sum_{i \in U} x_i^* \)” (Särndal & Lundström, 2005).

In the case of estimation under non-response Särndal and Lundström (2005) refer to the corresponding estimator over the response set:

\[ \hat{y}_W = \sum_{i \in r} d_i v_i y_i = \sum_{i \in r} w_i y_i \] 2-42

as the calibration estimator.

2.8 Reporting academic outcomes in educational surveys

A feature of many educational surveys including TIMSS, PISA and the Australian national surveys discussed in this thesis is that student achievement outcomes are
generally reported not as a single score but as a set of five so-called ‘plausible values’ or random draws from a posterior distribution used to estimate the student’s ability.

The measurement of student ability in mathematics for example will generally involve a mathematics test, but because the number of items that can be administered is limited and the limited coverage that any test can provide with respect to the vast field known as mathematics, there is considerable measurement error associated with any estimate of student ability from such a test. Whilst a point estimate of student ability derived from the test would be most appropriate at the individual level (for example in the comparison of percentage correct between two students who completed the same test), there is considerable measurement error attached to those individual point estimates.

In educational surveys the major outcomes of interest are not generally at the individual level but at the level of the population, for example the average student achievement for a State, or the proportion of students in the population who achieved a certain benchmark of achievement.

When making estimates at the population level, plausible values incorporate the measurement error component of the estimate which, along with sampling error, contributes to the standard error of the estimate. “Plausible values provide not only information about a student’s ability estimate, but also the uncertainty associated with this estimate” (Wu (2005)).

In the course of this thesis, the outcomes of interest will be estimates at the level of the population, and so plausible values will be used as the dependent variable.
2.9 Generalised regression estimation in an educational survey

Micklewright et al. (2012) make use of auxiliary data known for all pupils in England to develop generalised regression (GREG) estimators in their analysis of non-response bias in relation to England’s participation in the PISA 2000 and PISA 2003 surveys. The auxiliary data includes national assessment data not long before, or shortly after the time of the PISA survey. The outcomes of national examinations undertaken by students around the ages of 15 and 16 can be used in the analysis of outcomes on PISA, an assessment of students at approximately the same age. The authors note the high correlation (around 0.8) of the national achievement measures with the PISA outcomes. The authors also draw on other information, such as eligibility for free school meals as an indicator of the socio-economic background of the students.

Through the use of auxiliary data at the population level they generate estimates of the mean outcome quite different to those obtained when using weights that simply reflect the sample design or the non-response adjusted weights developed through the PISA survey. They are also make comparisons between GREG based weight adjustments making use of auxiliary data available at the population level and weights based on estimated response propensities (section 2.6.8) as a clear demonstration of the power of making use of auxiliary data available at the population level compared to being limited to sample and response level data derived during the survey.
2.10 Response propensity adjustments in a study of response bias in England

As noted in section 2.8, in addition to a GREG-based investigation, Micklewright, Schnepf and Skinner (2012) develop response propensity models using logistic regression in their analysis of response bias in the data collected for England in the 2000 and 2003 administrations of the PISA survey. In one model, the researchers estimate non-response as a function of gender, eligibility for free meals (as a measure of socio economic background of the students), achievement (based on national examination results), and private school attendance. Based on the expectation that response behaviour is likely to be in part determined by school characteristics, a dummy variable for each school was added to the regression model. Using this model they were able to derive a predicted probability of response based on these background factors and, by taking its inverse, an associated weight \( \hat{c}_i = \hat{\phi}_i^{-1} \). They then combined this weight with weights reflecting the sample design, as well as adjustments reflecting non-response at the school level to derive an overall student weight.

They contrast this approach with the weighting used in PISA, which involved the construction of weighting classes based on sample design variables, with school level factors involving school authority, prior performance and school sex composition. In particular they point out that no non-response adjustment was made by prior performance at the student level: “For example in a school where 35 pupils were sampled but only 25 responded, the OECD student weight would include a factor equal to 35/25”. 
Their analysis indicates that at least in the context of the England, the weighting adjustments that simply reflect school and student non-response primarily on the basis of school level sample design variables without adjusting for variations in student factors such as prior performance did not appear to substantially reduce the non-response bias. The weights they constructed that take response propensity into account lead to estimates of mean and standard deviation much closer to those observed for all sampled students.

As noted in section 2.8, one limitation with response propensity estimation as an adjustment technique is that it is limited to the use of sample data, it does not make use of auxiliary data at the population level. In relation to their analysis of response bias in the PISA data, Micklewright et al. (2012) show that while the response propensity adjusted weights, making use of the powerful auxiliary data available through England’s national assessments, performed much better than the PISA non-response adjusted weights, it was able to more accurately capture the effects of student-level non-response whereas the GREG-based approach, making use of auxiliary data at the population level was in addition able to explain non-response at the school level.

2.11 Study comparing properties of estimators

Särndal et al. (1992) conducted a simulation study testing the bias resistance properties of three different estimators – a simple weighting class estimator not making use of auxiliary information, and two estimators incorporating an auxiliary variable highly
correlated with the outcome variable, a ratio and a regression estimator. The empirical results over 1000 samples showed that the proportion of 95% confidence intervals covering the population total met or came very close to that figure for the estimators using the auxiliary information, even when the response model was quite wrong. However, when the data was erroneously assumed to be MCAR, the weighting class estimator performed very poorly.

As well as needing to identify a model that describes the response mechanism as accurately as possible, it is also essential that auxiliary information be strongly correlated with the outcome variables be collected to minimise the risk of bias.

Särndal et al. (1992) suggest three principles for selecting the vector of auxiliary variables. The variables selected should: (i) explain variation in response probabilities; (ii) explain variation of the main study variables; and (iii) identify the most important domains. They write: “The better we succeed in incorporating relevant auxiliary information into the [vector of auxiliary variables], the better, generally speaking, are the chances of realising a low nonresponse bias.”

2.12Chapter Summary

The review of the literature in the first part of this chapter summarised the development of approaches to the management of non-response over recent decades. This included a review of approaches undertaken in the preparation stages to maximise participation in the survey, and at the sample design stage, for example to ensure that important auxiliary information is collected in anticipation of non-response to be incorporated into estimation methods. A range of estimation methods were described, along with the
explicit or implicit assumptions underlying those methods, for example that sub-classes of the population could be identified within which an equal probability of response could be assumed. The methods extended to more complex approaches where auxiliary information was incorporated into estimators and into models in an effort to provide greater protection against biases arising from non-response, whilst maintaining precision in estimation.

Following the review of the literature, the formulation of a number of the approaches to be used within the later chapters of the thesis were presented to better illustrate how the methods include additional components associated with non-response and the relationships to precision and bias. The chapter concludes with an illustration of the application of key approaches to non-response management in research studies, particularly research that was conducted into non-response bias in relation to England’s participation in PISA 2003.

The following chapters will involve a study of non-response in relation to Australian educational surveys, and draw on the methods described in this chapter to investigate whether some of the methods might improve estimation for these surveys with respect to precision and bias, particularly with the inclusion of a wider range of auxiliary information, such as the data available from NAPLAN and MySchool.
Chapter 3  Explaining variation in educational achievement

3.1 Introduction

The review of methods for managing non-response in the previous chapter identified that approaches that incorporate information correlated with survey outcomes and/or with survey response can be very beneficial both with respect to protecting from the risk of non-response bias, and also for improving the precision of estimates. Chapter 3 examines research into educational outcomes from recent Australian survey, with the aim of identifying potentially useful auxiliary information to assist with non-response management approaches investigated in the later chapters.

As with any aspect of human development, the factors that result in differing levels of educational attainment are many and complex. Among recent papers, a meta-analysis of empirical findings from a large number of research papers was conducted by Hattie.
who presents findings across analyses including individual, family and school background factors. Karbach, Gottschling, Spengler, Hegewald, and Spinath (2013) discuss the relationship between school achievement and general cognitive ability, as well as parental involvement. The complexities associated with a particular individual’s educational outcomes, or with the pattern of outcomes from that person’s class or school are likely to extend beyond those able to be identified within a large scale survey. Nevertheless, there are broad influences that are shown across many surveys to be associated with educational outcomes. For example, socio-economic background has repeatedly been shown to be an important factor in student achievement. Similarly, aspects such as geographic location and access to resources consistently come up as variables that contribute to achievement, see for example Sirin (2005), Betts, Reuben, and Danenberg (2000).

This chapter will begin with a description of Australia’s National Assessment Program (NAP) which encompasses most of the survey activity being directed by the Commonwealth, State and Territory governments of Australia towards improving educational outcomes for Australia’s school students. The background to the NAP and the suite of surveys that are encompassed under this program will be described in section 3.2. Particular attention will be paid to the NAPLAN census-based survey and MySchool, a website designed to provide the community with information about the socio-educational and performance profile of Australian schools and to enable comparisons between schools with similar profiles. A key component facilitating those comparisons, the Index of Community Socio-educational Advantage (ICSEA), will be described.
NAPLAN and MySchool represent potentially important sources of auxiliary data which may be useful for managing non-response for other Australian educational surveys, but are currently used very little or not at all for this purpose. The primary research question for this thesis (section 1.3) is to investigate what improvements might be expected if such data were more widely incorporated into non-response management.

Following this description will be a review of recent studies examining the factors related to achievement using data generated from NAP surveys. These studies will explore the factors underlying achievement at various stages of schooling making use of NAPLAN data and observations from two international surveys – Trends in International Mathematics and Science Study (TIMSS), and Programme for International Student Achievement (PISA) – and also data external to the NAP, for example national tertiary entrance rank (ENTER) scores.

3.2 Australian student surveys – a recent history and background

In 1999 Australian Commonwealth, State and Territory ministers responsible for school education jointly agreed to a set of national goals for schooling. The preamble to the Adelaide Declaration on National Goals for Schooling in the Twenty First Century states: “Australia's future depends upon each citizen having the necessary knowledge, understanding, skills and values for a productive and rewarding life in an educated, just and open society. High quality schooling is central to achieving this vision.” Council (1999). Seventeen goals were listed under three broad categories, one category relating to the talents and capabilities of students, one specifically related to the curriculum, and
one relating to social justice. Since the publication of the Adelaide Declaration, a considerable amount of activity has been directed towards these national goals.

In 2008, the Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) published a *Measurement Framework for National Key Performance Measures (2008)*. The aim of this framework is described in the introduction as “driving school improvement and enhanced outcomes for students”, including the use of data collections “to improve Australian educational policy” ([Ministerial Council on Education (2008)]). The framework identifies a subset of the national goals to be prioritised, namely literacy; numeracy; science; civics and citizenship education. Policy towards these outcomes would be shaped through evidence collected from surveys of students. Since 2013, the ministerial representative body directing the NAP activities has been the Education Council:

*The ... Education Council provides a forum through which strategic policy on school education, early childhood and higher education can be coordinated at the national level and through which information can be shared, and resources used collaboratively, to address issues of national significance (Education Council (2017)).*

The program of assessments directed toward the national goals is referred to as the National Assessment Program (NAP). Table 3-1 summarises the assessment-based achievement surveys falling under the NAP framework since 2011. It comprises census testing - all students in Years 3, 5, 7 and 9 - in literacy and numeracy, conducted since 2008 on a national level, participation in major international comparative assessments, and a rotating schedule of national sample surveys in the national goal priority areas.
The administration and management of the NAP surveys, as well as the production of reports derived from these surveys, is the responsibility of the Australian Curriculum, Assessment and Reporting Authority (ACARA), an independent statutory authority directed by the Education Council.

Table 3-1: National Assessment Program surveys 2011-2017

<table>
<thead>
<tr>
<th>Survey</th>
<th>Abbreviation</th>
<th>Scope</th>
<th>Major assessment focus</th>
<th>Cohort</th>
<th>Survey design</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Literacy and numeracy benchmarks</td>
<td>NAPLAN</td>
<td>National</td>
<td>Reading, Writing, Language Conventions, Numeracy</td>
<td>Years 3, 5, 7, 9</td>
<td>Census</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Trends in International Mathematics and Science Study</td>
<td>TIMSS</td>
<td>International</td>
<td>Mathematics and Science</td>
<td>Years 4 &amp; 8</td>
<td>National sample</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Progress in International Reading Literacy Study</td>
<td>PIRLS</td>
<td>International</td>
<td>Reading</td>
<td>Year 4</td>
<td>National sample</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Programme for International Student Assessment</td>
<td>PISA</td>
<td>International</td>
<td>Reading, Mathematics, Science</td>
<td>15 year olds</td>
<td>National sample</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAP Science Literacy</td>
<td>NAP SL</td>
<td>National</td>
<td>Science</td>
<td>Year 6</td>
<td>National sample</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAP Civics and Citizenship</td>
<td>NAP CC</td>
<td>National</td>
<td>Civics</td>
<td>Years 6 and 10</td>
<td>National sample</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAP Information and Communication Technology Literacy</td>
<td>NAP ICTL</td>
<td>National</td>
<td>Information and Communication Technology</td>
<td>Years 6 and 10</td>
<td>National sample</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each of these surveys involves an extensive data collection activity. In addition to an assessment of the content domain, the surveys collect background information aimed at finding factors within each student’s learning context. The range of background data collected in TIMSS will be described in section 4.2.1.
3.2.1 NAPLAN and MySchool

The most significant of the NAP studies is NAPLAN, the census testing of students in Years 3, 5, 7 and 9 conducted annually since 2008. At each year level tests are administered in Language Conventions (40 minutes): Reading (between 45 and 65 minutes); Numeracy (between 45 and 60 minutes) and Writing (40 minutes). The Language Conventions tests includes items on Spelling, Grammar and Punctuation. The testing occurs over a three day period around May each year. Reports are prepared at all levels of participation, nationally, system level, school level and student level.

In a brochure prepared for parents and carers, the purpose of NAPLAN is described as follows:

*NAPLAN is the measure through which governments, education authorities, schools, teachers and parents can determine whether or not young Australians are meeting important educational outcomes in literacy and numeracy. The tests provide parents and schools with an understanding of how individual students are performing at the time of the tests. They also provide schools, States and Territories with information about how education programs are working and what areas need to be improved* (ACARA (2017e)).

School level outcomes from the NAPLAN assessment are a major contributor to the data published on the MySchool website managed by ACARA:

*The My School website is a resource for parents, educators and the community to receive important information about each of Australia’s schools in an easily accessible format.*

*My School contains data on such things as a school’s student profile, NAPLAN performance, funding levels and sources, and other financial information. You can also see enrolment numbers and attendance rates.* (ACARA ((2017a)))
3.2.2 MySchool comparisons across schools

One of the key functions of the MySchool website is to enable the comparison of schools with students from similar socio-educational backgrounds. The similarity of schools is determined by the Index of Community Socio-educational Advantage (ICSEA), an index developed primarily from data collected at the school level on student background characteristics.

The construction of the ICSEA scale has evolved since its first use in 2008. The NAPLAN data used in this thesis are from 2010, and the ICSEA construction in that year was made up for four components: socio-educational information, proportion of Aboriginal and Torres Strait Islander (ATSI) students; accessibility/remoteness; and the proportion of disadvantaged students from families with language backgrounds other than English (LBOTE). The socio-educational information was comprised of information relating to parent occupation, school education, non-school education and language background, usually obtained from student enrolment records, but in some cases data collected from the latest Australian Bureau of Statistics (ABS) census data linked to the Census District containing the home address (income, education and occupation, employment family composition) was preferred. Barnes (2011) provides details of the construction of ICSEA for 2010 and an evaluation of the respective data sources.

3.3 Previous studies explaining educational outcomes

There exists a very large literature exploring educational achievement and the home and school background factors that contribute to that achievement. For example in recent
years the OECD has sponsored a number of thematic reports drawing on the results across countries on the PISA survey. The latest publications, using data from the PISA 2015 survey, have included reports focussing on Excellence and Equity in Education; Policies and Practices for Successful Schools, Students’ Well-Being (OECD (2017)).

In the Australian context, reports of outcomes from national and international surveys have contributed over several decades to an understanding of the factors important in improving outcomes for students. Each major survey is accompanied with its own national report, for example Thomson, Hillman, and Wernert (2012a) in relation to Australia’s participation in the TIMSS 2011 survey, and Thomson, De Bortoli, and Buckley (2013) in relation to the PISA 2012 survey.

In addition, many researchers have used data from NAP studies to explore the relationship between school characteristics (including school level factors and averages of student level factors) and outcomes. Ainley and Gebhardt (2013) provide a meta-analytic overview of reporting of national and international surveys in which Australia has participated, extending back over more than 50 years. This report places achievement outcomes at different levels of schooling into broader changes in contextual factors such as an increasing national perspective on educational policy and governance, a greater emphasis on accountability in educational provision, changes in expenditure and changes to the distribution of enrolments across government and non-government school sectors. In a recent discussion paper noting a significant decline in reading and mathematical literacy levels of Australian 15 year olds since 2000 observed in the PISA study, Masters (2016) writes that
Most countries recognise that quality schooling and high levels of overall educational performance depend on reducing disparities between schools ... Not only is there evidence that Australia’s secondary schools became increasingly different over this period, but these performance disparities also became increasingly associated with average socioeconomic background.

A number of researchers have used NAPLAN data and other data such as tertiary entrance rankings to explore factors relating to educational outcomes. These include Marks (2010); Lamb, Rumberger, Jesson, and Teese (2004); Leigh (2010); Miller and Voon (2011), as discussed in more detail below.

3.3.1 Explaining NAPLAN outcomes at the school level

Miller and Voon (2011) made use of data available through the My School website to explore outcomes at the school level across States and types of schools across Australia. Examining Year 3 NAPLAN outcomes from the 2009 assessment, they found a collection of variables explaining variations in average school achievement on the NAPLAN assessments. In 2009 NAPLAN assessments were conducted at Year 3 in the domains of Reading, Writing, Spelling, Grammar and Numeracy.

Among 18 different explanatory variables investigated in a series of regression analyses, the authors found that ICSEA was by far the largest contributor to explaining variation in outcomes across schools. While ICSEA was the dominant explanatory factor, many other variables also contributed. Other variables studied included the size of the school, the proportion of female enrolments, dummies for the State, sector and location of the school, the type of school (primary, secondary or combined) and the attendance rate. A total adjusted $R^2$ of 63% was obtained when school average grammar scores were
regressed against ICSEA and these other variables. While ICSEA was the most important variable, more than 15% of the explanatory power was coming from the variables in addition to ICSEA.

Among the additional variables, important factors included the school composition by sex. The authors found, for example, that all other things equal, all-girls schools performed on average more half a standard deviation better than all-boys schools on the grammar test. School sector was another important variable with improved outcomes on the grammar assessment observed for Independent and Catholic schools compared to State schools. Differences were also observed between States and Territories. The authors attribute some of these differences to lower school entry ages, particularly for Queensland, where Year 3 students have one less year of schooling than their counterparts in other States. Positive coefficients were recorded for provincial, remote and very remote schools when referenced against schools from metropolitan regions. The authors attribute this apparently anomalous finding to the inclusion of ICSEA and attendance rate into the model. Finally, the school percentage attendance rate was found to be an important additional variable with large differences on the grammar assessment for schools with very high attendance rates compared to those with lower rates of attendance.

Across the other NAPLAN assessment domains, the results were largely similar. Some differences in the amount of variance explained across different variables appeared to be related to the construction of the ICSEA index, which at that time incorporated a school performance measure constructed from the reading and numeracy tests. One other notable difference was that the variable proportion of females which was highly
statistically significant for the models for most domains, was not significant for the numeracy assessment.

For similar analyses across later year levels, the authors find similar general patterns of effects. Differences between States appear to play a much more minor role with respect to Year 9 outcomes than for the earlier year levels.

3.3.2 Aspects of school effectiveness

In their efforts to examine the effectiveness of schools themselves in producing better outcomes for students academically and otherwise, Lamb et al. (2004) analyse outcomes after controlling for differences in student background. They find considerable segregation of students on the basis of social and academic background:

*Independent schools accounted for about 19 percent of all Victorian Year 12 students in 2000. However, they enrolled over 40 percent of all students from the highest SES band – those in the highest quintile of SES – and over 35 percent of all students from the highest general achievement band.*

Moreover, they find that student segregation tends to intensify school level differences in outcomes. One of their observations is the effect of factors such as selectivity, scholarships, and travel to attend schools of choice. They noticed this in particular with mostly independent schools operating in low socio-economic areas performing well above the levels predicted after controlling for socio-economic background:

*The poor location of the student’s address lowers the average social level of the school, while the student’s ability raises the performance profile of the school above levels expected purely from social intake....By contrast, a school whose social prediction of success is low may perform more poorly if it tends to be a refuge for young people who are not accepted elsewhere in the local community.*
In order to investigate the effectiveness of schools themselves to affect educational outcomes the authors apply a range of multi-level regression analyses looking at factors explaining achievement in mathematics at junior secondary level (Year 8) using data from the Third International Mathematics and Science Study (TIMSS, 1995). These analyses were aimed at identifying significant factors at both school and student levels and how these combine to explain variation in achievement outcomes. A further component of the analysis was to examine the level – between and within schools – where variance in outcomes was most present, and in so doing distinguish between factors where the influence of the school appeared to extend beyond student level factors.

First, applying a model with no predictors, the authors confirm the earlier finding of a segregated school system, finding that one quarter of the variation in student mathematics achievement is explained at the school level – i.e. is due to differences in the schools that the students attend.

From this basis, school level factors – school size, socio-economic background, prior achievement and school sector (government / catholic/ independent) were added into the model to explain achievement and these factors explained 86% of the variation between schools that was observed in the model with no predictors. In other words, the component of the total variance that was between schools reduced by that amount with the addition of those school level factors.

Additional school level variables associated with the teaching resources of the school were added into the multi-level model. These included average age, qualification levels,
years in teaching, teacher satisfaction and style of teaching. These latter two variables were found to be statistically significant when added into the regression model, with more traditional styles of teaching leading to somewhat lower outcomes compared to more innovative teaching practices, all else equal. Higher teacher satisfaction levels were also found to contribute positively to outcomes. An additional 2% of the between school variation was explained by these additional factors, results in up to 88% of the variation at that level.

Incorporating student level variables into the model resulted in some reduction in the coefficients for the school level factors. For example, the inclusion of the SES measure at the student level was a clearly significant factor while the school level measure of socio-economic background, which had been a significant factor in the earlier model involving only school level factors was now no longer significant. The importance of the school mean achievement measure in explaining outcomes dropped somewhat, but was still a significant and important variable in the model, with an increase in the TIMSS outcomes of around 10 points for each one point increase in the school mean (compared to around 16 points in the earlier model). The teacher variables remained as significant predictors in the model, and mediating variables at the student level such as attitudes towards maths were also significant. Overall, approximately 23% of the variation in achievement within schools was explained with the addition of the student variables, whilst the proportion of explained variation between schools remained at about 88%.

The authors conclude that the highly significant outcomes for the school level achievement and teacher factors in explaining differences even with student level
variables included in the model suggest that schools with concentrations of students scoring highly with respect to these variables provide a platform from which schools can add to outcomes. “Like physical resources, pupils provide a resource which helps some schools organize their teaching and other programs in ways which help raise levels of achievement”.

3.3.3 Explaining achievement at the upper end of secondary schooling

Marks (2010) uses longitudinal data obtained from the 2003 PISA survey to explore the factors explaining achievement at the upper end of secondary schooling. PISA is a survey of 15 year olds, typically around Year 10 in Australian schools, and this study followed the progress of students participating in PISA for their remaining years of secondary schooling, to the end of Year 12. This survey involved a nationally representative sample of 12551 students from 321 schools. The students sampled for this survey were also invited to participate in a longitudinal component where they would be followed up for subsequent surveys in future years. Of the original sample, 10448 were successfully interviewed by telephone in late 2003.

Telephone interviews were conducted in each of the following years, with some attrition in survey responses experienced each year. By 2006, responses were obtained from 7772 of the original sample. A weight was constructed combining the PISA final student weight, reflecting the sample design and school and student non-response, and weights to compensate for the attrition of students in the longitudinal component (Rothman (2007)). The outcome measure used was based on students’ tertiary entrance rank, or ENTER. “ENTER scores are percentile ranks ranging from 30 to 99.95. A score of 90 means that the student’s tertiary entrance rank was higher than 90% of the cohort”
Marks (2010). Only those students who received an ENTER score were included in the analysis of outcomes, the 20% to 25% of students in the study who had left school prior to Year 12, or those who did not receive an ENTER score were not included.

Marks (2010) conducts an analysis of factors explaining outcomes at year 12. He presents the results of a sequence of regression models, with ENTER score as the response variable, starting with student-level predictors including sex, location, language background, family size and school type. In the first series of regression analyses performed in this study predictor variables were limited to student-level variables. While no school effects are included in the first series of models, standard errors are adjusted to take account of the clustering of students within schools. The initial model in this first series explores demographic factors including Gender, Location and family type. While a number of these factors were found to be statistically significant, the model based on these factors explained just 3% of variation in the ENTER scores. He then progressively adds variables of different types including SES, location and average achievement, and student attitudes to school, as well as their evaluation of teacher efficacy and the learning environment of their classroom.

The SES measure used in this analysis was based on a PISA construct, a measure of Economic, Social and Cultural status (ESCS). This index is built from a number of components including the highest of either the father’s or mother’s occupational status (coded to an international index), indices of educational and cultural resources and the number of books in the home. The inclusion of this variable increased the explained variance from 3% to 12%. A one standard deviation increase in ESCS was associated with an increase of 5.6 points on the ENTER score.
However, the effect of SES was substantially reduced with the addition of the prior performance measure, based on the student’s PISA test scores in 2003. The inclusion of this variable increased the explained variance in ENTER scores from 12% to 40%. A difference of one standard deviation on the PISA test score measure was associated with a difference of over 12 points on the ENTER scale. In contrast, once the prior academic was included as a control, a one standard deviation difference in the SES background measure was associated with a change in ENTER rank of just 2.4.

*The strong effect for PISA test score when controlling for the ESCS measure of socioeconomic background shows that the effect of student achievement on tertiary entrance performance cannot be attributed to socioeconomic background.* (Marks (2010))

A second series of regression models is then applied, examining school effects. The analyses are undertaken using PROC MIXED in SAS which is noted as “appropriate for the analysis of multilevel data” (Marks 2010). No mention is made about the treatment with respect to the application of weights at the school and student levels. The first model involves adding random school effects to the final model applied in the first series. The between school variance under this model reduces by approximately 50% (from 25% to 13%) indicating that about this proportion of the between school variance is due to differences between students within the schools. Further declines in the between school variance were observed when other school level factors were added to the model. A moderate effect is observed when the school level socio-economic background variable is added, with a one standard deviation difference in school SES associated with a change of 2.5 ENTER score points. However when the average school
A number of other school level factors, particularly as perceived by students themselves were found to be significant in explaining student level outcomes when added to the model. More positive attitudes to school were associated with better outcomes, whereas perceptions of a poor disciplinary climate were negatively associated with outcomes. An interesting result was that at the student level, academic press – the extent to which students perceived the school as pressing for higher academic outcomes – appeared as negatively associated with outcomes. Marks (2010) suggests that this counterintuitive outcome may arise with higher achievers not perceiving their classmates as particularly hard working or eager to achieve. In contrast, at the school level, this variable performed according to theoretical expectations and was positively correlated with outcomes.

3.3.4 Summary of results of previous studies

The results of these previous research papers reveal that while the factors explaining educational outcomes are complex, some factors dominate with respect to their capacity to explain achievement outcomes – in particular the profile of the student group with respect to achievement (as measured for example by the average score achieved by students participating on the PISA survey), and student level socio-economic background factors. While across many studies a strong relationship exists between performance measure is added, the effect of this variable is of a similar order (2.6 ENTER score points) and the school level SES measure ceases to be a statistically significant factor. Marks (2010) concludes that “it is not the case that students’ tertiary entrance performance is influenced by the socio-economic context of the school but by its academic context”.

Marks (2010) concludes that “it is not the case that students’ tertiary entrance performance is influenced by the socio-economic context of the school but by its academic context”.

3.3.4 Summary of results of previous studies
The results of these previous research papers reveal that while the factors explaining educational outcomes are complex, some factors dominate with respect to their capacity to explain achievement outcomes – in particular the profile of the student group with respect to achievement (as measured for example by the average score achieved by students participating on the PISA survey), and student level socio-economic background factors. While across many studies a strong relationship exists between

Marks (2010) concludes that “it is not the case that students’ tertiary entrance performance is influenced by the socio-economic context of the school but by its academic context”.

3.3.4 Summary of results of previous studies
The results of these previous research papers reveal that while the factors explaining educational outcomes are complex, some factors dominate with respect to their capacity to explain achievement outcomes – in particular the profile of the student group with respect to achievement (as measured for example by the average score achieved by students participating on the PISA survey), and student level socio-economic background factors. While across many studies a strong relationship exists between
socio-economic background and outcomes, many schools perform either above or below what would be predicted by their socio-economic composition alone. Other factors such as teacher experience, disciplinary climate and school academic press each contribute significantly, though to a smaller degree in explaining variation in academic outcomes. It appears that having a better than average profile with respect to achievement and social factors enables many schools to establish school climate factors such as academic press and disciplinary climate to further enhance outcomes, in contrast, many schools with lower than average profiles on these factors have difficulty in overcoming these difficulties in their efforts with respect to the outcomes of their students.

3.4 Chapter summary and links to following chapters

Chapter 3 has summarised the extensive educational survey program that exists within Australia and its link to policy work extending back over more than twenty years. In particular it has paid attention to Australia’s National Assessment Program (NAP), and the central survey components of NAPLAN and MySchool. It has also examined recent research work undertaken that makes use of these important surveys and data. A key finding from this chapter is that some factors – particularly prior academic performance and socio-economic background - are clearly important in explaining educational achievement outcomes. It is also clear that the interaction between academic and socio-economic background and the mix of individual and school level effects with respect to these characteristics in their contribution to explaining achievement outcomes is complex. A number of other factors such as school climate factors were also found to be
significant contributors, although their contribution to explaining variance in achievement, when academic and socio-economic background factors were included was less.

Based on these findings, the following chapters will explore the potential value of making use of these factors towards the management of non-response in educational surveys, following the steps outlined below.

3.4.1 Exploratory data analysis and preparation for model building

Chapter 4 will involve an analysis of outcomes from the 2011 Trends in International Mathematics and Science (TIMSS) assessment at Year 8. Following a detailed summary of the TIMSS survey instrumentation, design and weighting, the chapter will summarise the outcomes of an exploratory data analysis to explore the relationships between the educational achievement variable and available explanatory variables and assess the explanatory power those variables, expecting to confirm that these correspond more or less with the published literature surrounding recent other surveys.

Following the exploratory data analysis, a series of regression models will be constructed which successively add variables from a range of data sources, examining the amount of variation explained with each successive model.

The school data used in these analyses of TIMSS outcomes will be divided into three categories, as follows:
• Those present within the sampling frame used in the selection of the school samples. Several of these variables will often be incorporated into the sample design, for example for stratification purposes.

• Data available about participating schools not present on the sampling frame but that could reasonably be drawn upon if found to be valuable – in particular for this research, data will be drawn from the publicly available information published on the MySchool website.

• Data generated through the TIMSS survey itself in its efforts to identify important background characteristics for explaining mathematics performance. These data will potentially include school, teacher and student level data.

By considering these categories of data, a picture will emerge of where data which may potentially be useful in understanding the factors that contribute to achievement are located and, for each category of data in turn, the potential additional value of the use of these data for non-response adjustment purposes, the subject of the later chapters of the thesis.
Chapter 4  The TIMSS survey and factors explaining achievement

4.1 Introduction

This chapter will involve a more detailed study of one of the NAP surveys, the Trends in International Mathematics and Science Study (TIMSS). Following a description of the background of the study and its aims and objectives, a detailed description of the approach to non-response adjustment applied in the 2011 survey, including school substitution and non-response adjustment weighting, will be provided.

In section 4.4, an analysis of data collected from Australian students participating in this study will be performed in order to look at the factors most related to achievement which might be used in estimation to reduce non-response bias. The initial investigations will examine achievement in mathematics – a student level variable – and how variables at the school level contribute to explaining this achievement. The restriction to school level variables in this section reflects current practice with respect
to non-response management for Australian surveys. The variables available for the non-response management of NAP surveys, for example, have been restricted to the use of school level variables available on the database used for sampling and incorporated into the sample design. These variables include State; sector; location; school size; school type (secondary enrolments only; or combined primary and secondary); and the proportion of Aboriginal and Torres Strait Islander (ATSI) students.

The initial investigation will be supplemented with the incorporation of the following school level data from MySchool:

- school mean prior performance;
- the ICSEA index of socio-educational advantage;
- the proportion of students with language backgrounds other than English (LBOTE);
- the proportion of ATSI students;
- the school attendance rate; and
- reported recurrent funding from government and private sources.

Following this analyses, some student level data will be incorporated into the models explaining achievement. This part of the investigation is intended to investigate what improvements in explanatory power might be possible if data such as student NAPLAN performance data and student level socio-economic measures were available for use in non-response management. Those actual data were not made available for the purposes of this research, and so the investigations in this chapter will use student-level data collected from TIMSS itself. For example, in the investigation of achievement in mathematics (one student level academic outcome of the TIMSS survey), this chapter
will use student achievement in science (another TIMSS outcome) as an explanatory variable as a proxy for a correlated NAPLAN performance measure. The measure of ‘Home Educational Resources’ (HER) constructed within TIMSS from student questionnaire data (see section 4.4.6) will be used as a student-level socio-economic background measure.

A further extension will be the use of additional background variables at both school and student levels, for example estimates of students’ confidence with mathematics and estimates of the academic climate of the school as indicated by the school principal. As with the academic and socio-economic background variables described above, these additional variables will be derived from the TIMSS survey itself. In practice, such data collected from a survey would not be available for non-response management for that actual survey. However, similar data are collected for Australian students through NAPLAN and other related NAP activities. The TIMSS derived estimates used in this investigation are intended as proxies for those data that are collected outside of TIMSS and indicate the potential explanatory power of such variables.

As with all surveys conducted under the NAP, it will be observed that very high rates of school participation and high rates of student participation are achieved in TIMSS. There is a significant national effort to achieve high rates of participation for TIMSS, as for the other NAP surveys. Schools are essentially mandated to participate in these surveys, and there is a considerable effort to work with schools to achieve high rates of participation within schools, including well-targeted information to parents communicating the importance of these surveys and the provision of follow up sessions if response rates are low.
NAP surveys do experience some non-response, and the question of how to manage that non-response applies to these surveys as for others. This will be examined in detail with respect to the practices of the TIMSS survey (section 4.3).

For student surveys that are conducted in Australian schools that fall outside the NAP, lower rates of participation at both the school and student levels are generally observed. These surveys may have most to gain from lessons learned with respect to non-response management arising from the investigation of the surveys conducted under the NAP.

4.2 The 2011 TIMSS survey

The 2011 Trends in International Mathematics and Science Study (TIMSS 2011) compared mathematics and science teaching and learning practices and outcomes across 45 countries and a further 14 sub-national entities (known as benchmarking participants). TIMSS was first conducted in 1995, with subsequent surveys every four years. Australia has participated in all six TIMSS surveys between 1995 and 2015 (Ainley and Gebhardt, 2013). TIMSS has two target populations, focussing on achievements at the fourth and eighth years of formal schooling respectively. In Australia these are Grade 4 and Year 8. At Grade 4, the student sample selected for TIMSS was also used for the Progress in International Reading Literacy (PIRLS). The focus for this chapter will be on the TIMSS survey of Year 8 students.

PIRLS and TIMSS are conducted jointly on behalf of the International Association for the Evaluation of Educational Achievement (IEA). The IEA is “an independent, international cooperative of national research institutions and governmental research
agencies. It conducts large-scale comparative studies of educational achievement and other aspects of education” (IEA Homepage, 2016). The TIMSS and PIRLS International Study Centre is based at the Lynch School of Education, Boston College. Australia’s participation in TIMSS 2011 was funded by the Australian, State and Territory governments as part of the NAP, and was managed by the Australian Council for Educational Research (ACER).

The aim of the TIMSS survey is to provide participating countries with both performance data and system level, school and student background data in the domain areas of mathematics and science that are comparable with other participating countries. Participating countries obtain perspectives on the relative strengths and weaknesses of their educational programs compared to other participants.

TIMSS and PIRLS share very similar methodologies and produce similar outputs with respect to reporting and data. Each collect and report internationally comparable data in the respective content domains, (i.e. Reading, Mathematics and Science) and also extensive school and student background data drawn from Student, School and Teacher questionnaire instruments. These data include insights into home background, school and home educational contexts, student attitudes, and teacher backgrounds.

An extensive online technical report detailing survey methods and procedures is provided following the conclusion of each survey, for example the documentation in relation to the 2011 TIMSS and PIRLS surveys, Martin and Mullis (2012)

4.2.1 Instrumentation

The TIMSS data is collected through the following data instruments:
• A cognitive assessment of approximately 80 minutes duration;
• A Student Questionnaire focussed on students family background, aspects of learning and learning contexts;
• A Parent Questionnaire (for the fourth grade survey) focussed on home background, early learning experiences and parental occupation, attitudes and experience;
• A Teacher Questionnaire focussed on qualifications, teaching practices, and classroom climate
• The School Questionnaire, answered by the principal, focussed on school resources, policies and practices, and on school climate.

4.2.2 Sample Design

Target Population

The international target population for TIMSS is students in their eighth year of formal schooling, Year 8 for Australia.

School Level Exclusions

School level exclusions were permitted on the basis of geographical inaccessibility, very small size (fewer than five students in the target grade); a very different curriculum structure from the mainstream educational system (for example hospital schools); or schools whose students all fit within the category of student level exclusions (described below). Geographic inaccessibility was determined using a geolocation measure developed specifically with respect to the provision of education services (see Jones (2004)), developed on behalf of the Commonwealth, State and Territory Education ministries, and used for comparative reporting for all surveys conducted under Australia’s NAP.
Student Level Exclusions

The within-school exclusions were specified as: students with functional disabilities; students with intellectual disabilities; and non-native language speakers.

Reported exclusion rates

Table 4-1 presents the reported exclusion rates for Australia’s participation in TIMSS at Year 8: (Thomson, Hillman, Wernert, et al. (2012)).

Table 4-1: Exclusion rates TIMSS 2011 Year 8 Australia

<table>
<thead>
<tr>
<th>School level exclusions</th>
<th>Within-school exclusions</th>
<th>Overall exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.30%</td>
<td>1.90%</td>
<td>3.20%</td>
</tr>
</tbody>
</table>

Rates of exclusion were similar at both school and student levels. The overall rate of student exclusions was well within the 5% specified in the TIMSS documentation (M. Joncas & Foy, 2012) to meet international standards of participation.

Approach to sampling

TIMSS employs a two-stage random sample design:

- At the first stage a sample of schools was drawn with probability proportional to the size (PPS) of the enrolment at the target grade (Year 8). The enrolment size data was obtained from the sampling frame, which contained enrolment information provided by the respective Commonwealth and State educational departments. The measure of size was therefore based on enrolment data for the year level collected in the school year previous to the year that the surveys were conducted.
For the States of Australia, one intact class was sampled from each sampled school at the target grade, and all students from the sampled class were included in the sample. In ACT and NT two intact classes were sampled from each sampled school. This was done for a number of reasons, including improving the yield of the survey in these small jurisdictions and mitigating against the potential for further clustering of students within classes of the same school. For example if there was streaming of classes by ability, the random selection of two classes would provide a better spread of student abilities than from a single class.

Under the assumption that the number of classes at a year level was proportional to the enrolment size at that year level, the combination of sampling schools with PPS and the equal probability selection of classes or class units from the sampled school ensure that all students from the stratum had approximately equal probability of selection into the sample.

**Explicit Stratification**

The TIMSS samples were explicitly stratified by jurisdiction (State/Territory). Separate, independent samples of a fixed number of schools were drawn from each jurisdiction.

**Implicit stratification**

Within each explicit stratum the sample frames were sorted by geographic location (metropolitan, provincial and remote); socio economic background deciles based on the postcode of the school, and Sector (Catholic, government and independent).
Within strata, socio-economic background deciles were formed from the Australian Bureau of Statistics (ABS) Socio-Economic Indices for Areas (SEIFA) Education and Occupation Index (IEO) ABS (2011). The postcode of the school was linked to the ABS postal area and the corresponding SEIFA measure for this area was used.

The combination of sorting of the explicit strata by geographic location, IEO and sector and the systematic PPS sample selection ensured that the selected sample within each explicit stratum was proportional to the population with respect to these variables. In other words, the sample was implicitly stratified by these variables.

**Oversampling of smaller States and Territories**

An objective of the survey was to have sufficient data for reliable comparisons between jurisdictions. Schools and students from the smaller jurisdictions were sampled at a higher rate so that sufficient data was collected for reliable estimates of outcomes at this level. When data were aggregated across States and Territories, weights were required to ensure that participating students contributed to survey estimates according to the number of students in the population each participating student was representing.

**Substitute schools**

As for all of the surveys conducted under the NAP, a measure aimed at reducing the potential negative impact of non-response bias in TIMSS was the use of substitute schools in cases where the sampled school does not participate in the survey. At the time of sampling, up to two schools are selected as possible replacements for each sampled school. The schools assigned as replacement schools are those adjacent to the sampled school on the frame. Due to the organisation of the sampling frame prior to sampling described above replacement schools generally match the sampled school with
respect to the major stratification variables, i.e. (within explicit strata defined by State and Territory): sector, geographic location, IEO decile and size.

In recent years, participation in NAP surveys has been tied to Commonwealth funding arrangements, with the result that participation has become effectively mandatory. The unweighted school participation rate for both PIRLS Grade 4 and TIMSS Year 8 was 95%. Nevertheless, for various reasons the employment of replacement schools is adopted in a very small number of cases.

**State/Territory Sample sizes**

Table 4-2 shows the number of schools sampled for TIMSS Year 8, the number of those schools found to be ineligible (for example, schools that had recently closed), the number of originally sampled schools that participated, the number of replacement schools used, and the number of non-participating schools, (‘refusal schools’).

<table>
<thead>
<tr>
<th>Explicit Stratum (State/Territory)</th>
<th>Total Sampled Schools</th>
<th>Ineligible Schools</th>
<th>Participating Schools</th>
<th>First Replacement</th>
<th>Second Replacement</th>
<th>Refusal Schools</th>
<th>Excluded Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian Capital Territory</td>
<td>30</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New South Wales</td>
<td>45</td>
<td>1</td>
<td>43</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Northern Territory</td>
<td>15</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Queensland</td>
<td>45</td>
<td>1</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>South Australia</td>
<td>40</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Tasmania</td>
<td>30</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Victoria</td>
<td>45</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Western Australia</td>
<td>40</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>290</strong></td>
<td><strong>3</strong></td>
<td><strong>276</strong></td>
<td><strong>0</strong></td>
<td><strong>1</strong></td>
<td><strong>10</strong></td>
<td><strong>0</strong></td>
</tr>
</tbody>
</table>

Table 4-3 shows the achieved number of Year 8 schools and students by State and Territory. Also the weighted number of students is presented, which is an estimate of the student population size that the sampled students are representing. This is also given
for each State and Territory as a percentage of the total estimated population size for Australia. Weighting for the TIMSS survey will be discussed in Section 4.3.

Table 4-3: Student participation by State and Territory, TIMSS 2011 Year 8

<table>
<thead>
<tr>
<th>State / Territory</th>
<th>N (schools)</th>
<th>N (students)</th>
<th>Weighted N students</th>
<th>Weighted students (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian Capital Territory</td>
<td>ACT</td>
<td>30</td>
<td>1302</td>
<td>4961</td>
</tr>
<tr>
<td>New South Wales</td>
<td>NSW</td>
<td>43</td>
<td>1134</td>
<td>84570</td>
</tr>
<tr>
<td>Northern Territory</td>
<td>NT</td>
<td>10</td>
<td>452</td>
<td>2297</td>
</tr>
<tr>
<td>Queensland</td>
<td>QLD</td>
<td>43</td>
<td>1198</td>
<td>52199</td>
</tr>
<tr>
<td>South Australia</td>
<td>SA</td>
<td>39</td>
<td>888</td>
<td>18792</td>
</tr>
<tr>
<td>Tasmania</td>
<td>TAS</td>
<td>30</td>
<td>752</td>
<td>6691</td>
</tr>
<tr>
<td>Victoria</td>
<td>VIC</td>
<td>43</td>
<td>958</td>
<td>65361</td>
</tr>
<tr>
<td>Western Australia</td>
<td>WA</td>
<td>38</td>
<td>872</td>
<td>17114</td>
</tr>
<tr>
<td>Total</td>
<td>276</td>
<td>7556</td>
<td>251985</td>
<td>100</td>
</tr>
</tbody>
</table>

Additional Indigenous student sample

In order to boost the number of Indigenous students included in each survey, all Indigenous students from the sampled school, including those from classes other than those sampled, were included into the sample

Overlap control

Samples were drawn simultaneously at both Grade 4 (TIMSS and PIRLS) and Year 8 (TIMSS). Statistical overlap control procedures (OECD (2006)) were implemented to ensure that there was no overlap between the Grade 4 and Year 8 school samples. This was done to minimise the burden of participation for individual schools. The process involves selecting the sample at one year level, and then modifying the selection probabilities of schools selected for that sample prior to the selection of the sample for the second year level.
4.3 School and Student weight calculations for TIMSS

A weight is calculated for each student in the TIMSS database, which provides the relative contribution of each student to outcomes. The final student weight is made up of several factors. One set of factors reflect the probabilities of selection of schools, classes and students respectively. These are called ‘design weights’ or ‘base weights’. At each stage of sampling an adjustment is made to adjust for non-response that has occurred at that stage – the non-response of schools, classes within sampled schools, or students within sampled classes. The final weight is the product of these factors. A fuller description of each factor is presented below. The description is drawn from Joncas and Foy (2012).

4.3.1 School level weights:

Notation

- Within the population are strata h, h = 1…H.
- Within a stratum will be schools i, i = 1…I, and within those schools classes j, j=1…J.
- The number of elements sampled from the ith stratum, n_h.
- The number of classes sampled from the ith school, c_i.
- The total measure of size for the stratum, M_h. The measure of size of the ith school from stratum h, m_i,h.
- Superscripts sc cl st refer respectively to schools, classes and students.
- Subscripts s, r1, r2, nr refer respectively to sampled, first replacement, second replacement and non-responding schools.
- Superscripts rs and nr denote respectively responding students and non-respondents.
• Weight components reflecting sample probabilities (‘base’ or ‘design’ weights) will be denoted $BW$. $BW_{i,h}^{sc}$ denotes the base weight for the $i^{th}$ school from stratum $h$.

• Adjustments to the base weights will be denoted with $A$. $A_{i,h}^{cl}$ denotes a weighting adjustment made to the base weight of the classes sampled from the $i^{th}$ school from stratum $h$.

• The stratum subscript $h$ may be dropped when it is clear from the context. $A_{i,h}^{cl}$ may be reduced to $A_{i}^{cl}$ for example.

• The product of the base weights and weighting adjustments for a group will be denoted $FW$. $FW_{i,h}^{sc} = BW_{i,h}^{sc} * A_{i,h}^{sc}$

• The overall student sampling weight, the product of the base weights and adjustments at the school, class, and student levels – the same for all students in class $j$ of school $i$ - will be denoted $W_{i,j}$.

School design weight

The design weight for the $i^{th}$ School sampled from stratum $h$ is equal to the inverse of the probability of selection of that school:

$$BW_{i,h}^{sc} = \frac{M_h}{n_h \cdot m_{hi}}$$ \hspace{1cm} 4-1

$$M_h = \sum_{i=1}^{N_h} m_{hi}$$ \hspace{1cm} 4-2

Where $n_h$ is the number of sampled schools, $m_{hi}$ is the measure of size for the $i^{th}$ school from the stratum, $N_h$ is the total number of schools in stratum $h$ and $M_h$ is the total measure of size of those $N_h$ schools.
**School Non-participation adjustment**

A school-nonparticipation adjustment weight is calculated for each stratum as the number of eligible sampled schools from the stratum divided by the number of participants. Replacement schools were counted in the numerator of this factor. For readability, the stratum index, \( h \), is removed from all terms.

\[
A_{sc}^{h} = \frac{n_{s} + n_{r1} + n_{r2} + n_{nr}}{n_{s} + n_{r1} + n_{r2}}
\]

\( A_{sc}^{h} \) is the school non-participation adjustment for the explicit stratum; \( n_{s} \) is the number of sampled participating schools, \( n_{r1} \) the number of first replacement schools that participated, \( n_{r2} \) the number of second replacement schools that participated, \( n_{nr} \) the number of sampled schools that did not participate and were not replaced.

Sampled schools that were found to be ineligible were not replaced, and were not included in this calculation. For example this would include a school that was found to contain no students in the target population, a school that had recently closed, or a school such as a special school catering exclusively to students who would be excluded that was included on the frame for sampling by mistake.

**Final School Weight**

The final school weight for school \( i \) is simply the product of the school design weight and the school non-participation adjustment, calculated for each stratum separately:

\[
FW_{ih}^{sc} = A_{h}^{sc} \cdot BW_{ih}^{sc}
\]
4.3.2 Class weights

The weighting calculations within school reflected the two steps of sampling classes and then including all students from those sampled classes respectively, as described below.

*Classes design weight*

The within school design weight reflects the probability of selection of students from within each sampled school. In TIMSS, two classes were sampled from participating schools from ACT and NT, and one for the participating schools from the States.

The within school design weight corresponded to the total number of class units at the school divided by the number of sampled class units (usually one or two):

\[
BW_{i}^{cl} = \frac{C_i}{c_i}
\]

\(C_i\) is the total number of class units at sampled school \(i\), and \(c_i\) is the number of sampled class units.

*Class non-participation adjustment*

At the stratum level, class non-participation adjustments were generated to account for cases where a sampled class did not participate, or where the student participation within the sampled class was lower than 50 percent. The adjustments were applied at the stratum level rather than the school level to minimise the risk of bias. For example if classes of mathematics students at Year 8 were organised by ability, then a school level
class weighting adjustment might add bias by increasing the weights of an above- or below- average participating class.

As the risk of bias increases as participation rates decline, it was decided that a class with a participation rate lower than 50% would be treated as a non-responding class, and a zero weight would be assigned to the participating students from that class.

The class non-participation adjustment for stratum h was:

\[ A_{h}^{cl} = \frac{\sum_{i,h}^{} \frac{1}{s+r+1+r^2} \delta_i}{\sum_{i,h}^{} \frac{s+r+1+r^2}{c_i}} \]  

Where \( c_i \) is the number of sampled class units from the \( r^{th} \) school, and \( \delta_i \) is the number of class units that participated.

**Final Class weight**

The final class weight calculated for class \( j \) sampled from school \( i \) was the product of the class design weight and the class non-participation adjustment:

\[ FW_{i,j}^{cl} = A^{cl} \cdot BW_{i}^{cl} \]
4.3.3 Student Weights

*Student Design weight*

The student design weight was the inverse of the probability of selection of students from the sampled class. For Australia, all eligible students from the sampled class were included in the sample, and so the student design weight was therefore 1 for all students.

\[ BW_{ij}^{st} = 1 \]  

*Student non-participation adjustment*

The student non-participation adjustment for students sampled from class \( j \) of school \( i \) was calculated as the number of eligible students from that class divided by the number who participated.

\[ A_{ij}^{st} = \frac{s_{ij}^{rs} + s_{ij}^{nr}}{s_{ij}^{rs}} \]  

\( s_{ij}^{rs} \) is the number of participating students in the \( j^{th} \) class of the \( i^{th} \) school, and \( s_{ij}^{nr} \) is the number of eligible non-participating students.

*Final student level weight*

The final student level weight was the product of the design weight and the student non-participation adjustment, which for students participating in the Australian survey was equal to the non-participation adjustment.
\[ FW_{ij}^{st} = A_{ij}^{st}, BW_{ij}^{st} = A_{ij}^{st} \]

**Overall sampling weight**

The overall sampling weight was the product of the weights calculated at the school, class and student levels:

\[ W_{i,j} = FW_{i,j}^{sc}.FW_{i,j}^{cl}.FW_{i,j}^{st} \]

4.3.4 Non-response in additional background instrumentation and data

As noted in section 4.2.1, in addition to the information collected directly from students and their schools, information is also collected from teachers of the sampled students, and, for the fourth grade population, their parents. The TIMSS and PIRLS User Guide advise that these data should be merged with student level records in the conduct of any analyses. The variables from these instruments “... are in essence attributes of students and must be analysed in the same manner as student background variables.” (P. Foy (2013)). In addition to considerations with respect to school and student non-response, the usefulness of these background data depends on teacher and parent participation for these instruments. There are no additional weighting adjustments applied for the home questionnaire data. In Australia’s participation at year 4 in 2011 response rates for the parent questionnaire were reported as being between 50% and 70% (Mullis, Martin, Foy, and Arora (2012)), and estimated outcomes in the international report were flagged with respect to possible data quality issues. Whilst no parent questionnaire is administered at year 8, similar issues with respect to unit non-response and their effect
on data quality, particularly in relation to the teacher instrument are noted later in the thesis (section 4.4.6, Table 4-19).

4.3.5 Dependent variable
The outcome measure of primary interest in the analyses is the estimate of student ability in mathematics. Rather than a single point estimate of ability, student ability estimates in TIMSS (and all other major national and international educational surveys that Australia participates in) are represented by five plausible values, or random draws from a posterior distribution of ability that is estimated for the student. As described at section 2.8, this approach provides unbiased estimates of the population parameters such as the mean and variance and better takes account the measurement error associated with the assessment.

4.3.6 Variance estimation
For each estimate of student ability, a corresponding estimate of variance is required. Variance estimation is performed using Jacknife replication. 75 sample replicates are provided with the international database (Foy, Arora, and Stanco (2013)). SAS procedures provided with the international TIMSS database were used to perform the Jacknife variance estimation. The sample estimates and corresponding standard errors were generated for each of the five plausible values and then the outcomes averaged to obtain the final sample estimates.
4.4 Mathematics achievement outcomes

4.4.1 Outcomes from the Australian national report

The following summarises mathematics achievement outcomes as presented in the Australian national report Thomson, Hillman, and Wernert (2012b):

- Some differences were found between States and Territories. Students from the Australian Capital Territory (ACT) performed better on average in mathematics than all other jurisdictions except New South Wales (NSW). Students from NSW and Victoria (VIC) had better average outcomes than the other jurisdictions (besides ACT).

- The estimated mean mathematics performance was not statistically different between boys and girls nationally or within State or Territory.

- There were statistically significant differences between students according to the number of books they reported at home, and by the level of parental education.

- There was a large difference in outcomes between non-Indigenous and Indigenous students.

- The estimated achievement for students from non-English speaking backgrounds was higher than for those from English speaking backgrounds, but the difference was not statistically significant at the 5% level.

- The mean performance for students attending schools in metropolitan locations was statistically significantly higher than for those from remote locations.

- The mathematics outcome for students with more affluent than disadvantaged students was significantly higher than for schools with a more balanced socio-economic profile, as well as for those with a higher proportion of disadvantaged students.
4.4.2 Further sub-population comparisons

Further subpopulation comparisons in addition to those published in the national report are presented below. These include comparisons by school sex composition, school sector, indicators of socio-economic background, and prior performance.

Table 4-4 presents weighted estimates of mathematics achievement by the sex composition of the school, whether single sex or co-educational. Standard errors are estimated using Jackknife replication to reflect the complex sample design.

<table>
<thead>
<tr>
<th>Sex composition of the school</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-educational</td>
<td>6553</td>
<td>212668</td>
<td>80.1</td>
<td>495.1</td>
<td>4.8</td>
</tr>
<tr>
<td>Female</td>
<td>587</td>
<td>24594</td>
<td>88.5</td>
<td>546.5</td>
<td>14.6</td>
</tr>
<tr>
<td>Male</td>
<td>416</td>
<td>14723</td>
<td>99.8</td>
<td>575.3</td>
<td>39.1</td>
</tr>
<tr>
<td>Fullsample</td>
<td>7556</td>
<td>251985</td>
<td>85.4</td>
<td>504.8</td>
<td>5.1</td>
</tr>
</tbody>
</table>

The sample sizes for the single sex schools are small leading to large standard errors around the estimates. Nevertheless the estimated mean performance for students from single sex schools are higher than those from co-educational schools.

Table 4-5 presents outcomes by sector.

<table>
<thead>
<tr>
<th>subgroup</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catholic</td>
<td>1726</td>
<td>58021</td>
<td>74</td>
<td>519.4</td>
<td>8.8</td>
</tr>
<tr>
<td>Government</td>
<td>4566</td>
<td>150163</td>
<td>86</td>
<td>487.9</td>
<td>5.3</td>
</tr>
<tr>
<td>Independent</td>
<td>1264</td>
<td>43800</td>
<td>83</td>
<td>543.6</td>
<td>13.7</td>
</tr>
<tr>
<td>Full Sample</td>
<td>7556</td>
<td>251985</td>
<td>85</td>
<td>504.8</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Mean performance outcomes for students from the non-government sector are statistically significantly higher than for students from government schools.
The following table presents estimated outcomes by national deciles of the SEIFA Education and Occupation (IEO) index based on the location of the school. These data, available on the national sampling frame, were linked to the TIMSS data.

Table 4-6 Year 8 Mathematics outcomes by SEIFA IEO, TIMSS 2011

<table>
<thead>
<tr>
<th>SEIFA Deciles</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decile1</td>
<td>609</td>
<td>18918</td>
<td>78.9</td>
<td>444.9</td>
<td>13.6</td>
</tr>
<tr>
<td>Decile2</td>
<td>505</td>
<td>18252</td>
<td>69.1</td>
<td>459.3</td>
<td>12.9</td>
</tr>
<tr>
<td>Decile3</td>
<td>642</td>
<td>25164</td>
<td>73.9</td>
<td>484.2</td>
<td>9.8</td>
</tr>
<tr>
<td>Decile4</td>
<td>439</td>
<td>19735</td>
<td>74.2</td>
<td>476.9</td>
<td>10.9</td>
</tr>
<tr>
<td>Decile5</td>
<td>675</td>
<td>33841</td>
<td>72.9</td>
<td>486.2</td>
<td>11.6</td>
</tr>
<tr>
<td>Decile6</td>
<td>514</td>
<td>20066</td>
<td>68.3</td>
<td>511.5</td>
<td>9.6</td>
</tr>
<tr>
<td>Decile7</td>
<td>497</td>
<td>20343</td>
<td>76.2</td>
<td>517.1</td>
<td>16.4</td>
</tr>
<tr>
<td>Decile8</td>
<td>752</td>
<td>25384</td>
<td>71.4</td>
<td>490.2</td>
<td>8.6</td>
</tr>
<tr>
<td>Decile9</td>
<td>892</td>
<td>24224</td>
<td>80.0</td>
<td>539.2</td>
<td>16.5</td>
</tr>
<tr>
<td>Decile10</td>
<td>1796</td>
<td>44755</td>
<td>86.5</td>
<td>569.0</td>
<td>15.4</td>
</tr>
<tr>
<td>All records with seifa deciles</td>
<td>7321</td>
<td>250683</td>
<td>85.4</td>
<td>505.1</td>
<td>5.1</td>
</tr>
</tbody>
</table>

A clear relationship exists between the SEIFA deciles and the estimated mean mathematics outcome. Students from schools in the decile 1 have an estimated mean of 444.9, around 1.5 (overall) standard deviations lower than those in decile 10 (569.0).

A similar relationship is present in the estimated outcomes from one of the school questionnaire items about the average income of the school’s immediate area as reported by the school principal.

Table 4-7: Year 8 Mathematics outcomes by average income, TIMSS 2011

<table>
<thead>
<tr>
<th>average income</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>1091</td>
<td>34572</td>
<td>94.9</td>
<td>559.5</td>
<td>19.3</td>
</tr>
<tr>
<td>Medium</td>
<td>4189</td>
<td>143508</td>
<td>77.4</td>
<td>514.0</td>
<td>6.3</td>
</tr>
<tr>
<td>Low</td>
<td>1618</td>
<td>56952</td>
<td>75.6</td>
<td>461.0</td>
<td>6.9</td>
</tr>
</tbody>
</table>

As discussed earlier, the national report finds relationships for mathematics achievement with student background information about the number of books in the home and parental educational level. In addition, a scaled variable, the Home Educational Resources index was developed for TIMSS using responses from students.
to questions about three home resources: number of books in the home; number of
home study supports, and highest level of education of either parent (Foy et al. 2013).

The following table shows how outcomes vary across levels of this index averaged at
the school level. Once again, a clear relationship is shown between socio-economic
indicators and outcomes.

Table 4-8: Year 8 Mathematics outcomes by Home Education Resources index, TIMSS
2011

<table>
<thead>
<tr>
<th>Percentiles of School Mean Home Educational Resources</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>p10</td>
<td>629</td>
<td>22481</td>
<td>69.3</td>
<td>419.2</td>
<td>9.5</td>
</tr>
<tr>
<td>p25</td>
<td>1136</td>
<td>38638</td>
<td>65.9</td>
<td>462.2</td>
<td>6.2</td>
</tr>
<tr>
<td>p50</td>
<td>1787</td>
<td>68757</td>
<td>67.3</td>
<td>481.2</td>
<td>5.1</td>
</tr>
<tr>
<td>p75</td>
<td>1953</td>
<td>59131</td>
<td>65.3</td>
<td>521.0</td>
<td>6.6</td>
</tr>
<tr>
<td>p90</td>
<td>1192</td>
<td>37386</td>
<td>71.2</td>
<td>539.3</td>
<td>8.4</td>
</tr>
<tr>
<td>p90 plus</td>
<td>859</td>
<td>25590</td>
<td>73.0</td>
<td>620.0</td>
<td>14.3</td>
</tr>
</tbody>
</table>

The following tables present outcomes using data from MySchool that has been linked
to the TIMSS database. The first compares outcomes across levels of the 2010 ICSEA
score for the school, showing once again the relationship between outcomes and socio-
economic background.

Table 4-9: Year 8 Mathematics outcomes by ICSEA percentile groups, TIMSS 2011

<table>
<thead>
<tr>
<th>Percentiles of school ICSEA score</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>p10</td>
<td>911</td>
<td>22182</td>
<td>73.9</td>
<td>440.3</td>
<td>9.4</td>
</tr>
<tr>
<td>p25</td>
<td>1095</td>
<td>39340</td>
<td>80.0</td>
<td>470.4</td>
<td>10.5</td>
</tr>
<tr>
<td>p50</td>
<td>1657</td>
<td>61855</td>
<td>69.1</td>
<td>480.0</td>
<td>5.3</td>
</tr>
<tr>
<td>p75</td>
<td>1924</td>
<td>66993</td>
<td>68.4</td>
<td>508.7</td>
<td>6.9</td>
</tr>
<tr>
<td>p90</td>
<td>1262</td>
<td>37145</td>
<td>67.8</td>
<td>560.5</td>
<td>8.6</td>
</tr>
<tr>
<td>p90 plus</td>
<td>707</td>
<td>24470</td>
<td>97.3</td>
<td>586.2</td>
<td>24.9</td>
</tr>
</tbody>
</table>

Table 4-10 below compares outcomes by levels of the average Numeracy mean score
for Year 9 students using 2010 NAPLAN data. Not surprisingly, a strong relationship
exists between this measure of ‘prior-performance’ for the school and the TIMSS outcomes measures.

Table 4-10: Year 8 Mathematics outcomes by school mean numeracy score at Year 9, TIMSS 2011

<table>
<thead>
<tr>
<th>School Mean NAPLAN Numeracy score (year 9)</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>p10</td>
<td>839</td>
<td>18015</td>
<td>69.7</td>
<td>431.2</td>
<td>9.5</td>
</tr>
<tr>
<td>p25</td>
<td>1059</td>
<td>35740</td>
<td>71.1</td>
<td>463.0</td>
<td>8.7</td>
</tr>
<tr>
<td>p50</td>
<td>1817</td>
<td>60894</td>
<td>72.3</td>
<td>483.6</td>
<td>6.3</td>
</tr>
<tr>
<td>p75</td>
<td>1929</td>
<td>66184</td>
<td>69.0</td>
<td>501.6</td>
<td>6.8</td>
</tr>
<tr>
<td>p90</td>
<td>1190</td>
<td>39277</td>
<td>70.2</td>
<td>551.2</td>
<td>9.0</td>
</tr>
<tr>
<td>p90 plus</td>
<td>722</td>
<td>31874</td>
<td>92.6</td>
<td>583.3</td>
<td>21.9</td>
</tr>
<tr>
<td>Full Sample</td>
<td>7556</td>
<td>251985</td>
<td>85.4</td>
<td>504.8</td>
<td>5.1</td>
</tr>
</tbody>
</table>

The following table shows outcomes across different sized schools using data linked from the sampling frame.

Table 4-11: Year 8 Mathematics outcomes by school size groups, TIMSS 2011

<table>
<thead>
<tr>
<th>school size</th>
<th>n</th>
<th>n (wgt)</th>
<th>std. dev.</th>
<th>mean</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>p10</td>
<td>483</td>
<td>16447</td>
<td>74.1</td>
<td>493.2</td>
<td>11.2</td>
</tr>
<tr>
<td>p25</td>
<td>1138</td>
<td>38486</td>
<td>77.5</td>
<td>484.9</td>
<td>8.1</td>
</tr>
<tr>
<td>p50</td>
<td>1855</td>
<td>69362</td>
<td>96.4</td>
<td>525.5</td>
<td>11.1</td>
</tr>
<tr>
<td>p75</td>
<td>1956</td>
<td>64977</td>
<td>78.3</td>
<td>506.7</td>
<td>9.4</td>
</tr>
<tr>
<td>p90</td>
<td>1193</td>
<td>37128</td>
<td>84.2</td>
<td>489.5</td>
<td>13.5</td>
</tr>
<tr>
<td>p90 plus</td>
<td>931</td>
<td>25585</td>
<td>77.9</td>
<td>503.4</td>
<td>10.9</td>
</tr>
</tbody>
</table>

There is some indication that middle-sized schools tend to obtain better outcomes overall than larger or smaller schools, although the relationship is weak.

4.4.3 Multiple regression analyses

The descriptive analyses of section 4.4.2 illustrated that across a range of variables relating to the characteristics of Australian schools, differences in average mathematics outcomes are present. This section will examine how these variables work in combination to explain mathematics achievement outcomes.
A series of multiple regression models were examined. The dependent variable was the student level estimate of mathematics performance. Explanatory variables included sample design variables; variables from outside data sources; and variables collected as part of the TIMSS survey activity. Of primary interest for the analysis is the extent to which variables that to this point have not been available when decisions regarding non-response adjustments are made could improve the explanatory power of models which might be used for that purpose.

In the first instance, only variables currently available for potential use in non-response adjustments are included in the regression models. These are all school level variables. The outcomes from these investigations give a baseline indication of the explanatory power of currently available data.

Following that, school level variables collected through NAPLAN and other NAP activities and available on MySchool are added to the model. To date, these variables have not been made directly available for non-response management, estimation or analysis purposes for TIMSS or other NAP surveys. However, as these variables are made available to the public via MySchool, if the case can be made that these variables assist in improving the quality of non-response management and analysis of the survey being undertaken they could be made available for this purpose.

Following from the addition of the school level variables published on MySchool, a further extension would be the use of data collected at the student level. From the descriptive analysis discussed above, two obvious variable types to consider are scores from the NAPLAN assessments as measures of prior performance and a student level
indicator of socio-economic background. Following from the review of prior research into the factors relating to educational outcomes – for example Lamb et.al.’s (2004) observation that “Like physical resources, pupils provide a resource which helps some schools organize their teaching and other programs in ways which help raise levels of achievement”, further variables that capture aspects of school climate may also be valuable for this purpose.

While student level NAPLAN scores were not available for this research, the TIMSS database provides an additional student level performance measure, the student’s score on an assessment of science. For the purposes of this research, this will be used as a proxy for the student’s score on the NAPLAN assessment. As the mathematics and science assessments for TIMSS are conducted at the same time, and these assessments are drawn from the same assessment construct, the correlation between the two variables is high. When the maths outcome is predicted solely by the science outcome, the coefficient of determination ($R^2$) is 0.84. This may be higher than might be expected when using, for example, the Numeracy score achieved by the TIMSS students on the most recent NAPLAN assessment because of, for example, differences in the timing of the assessment in the school year, and differences in the assessment construct. Nevertheless as each is a carefully developed assessment of student ability in mathematics, one would also expect a high correlation between mathematics achievement measures estimated by NAPLAN and by TIMSS.

The analyses were conducted using PROC SURVEYREG in SAS. The school was identified as the primary sampling unit, and the explicit and implicit stratum variables were specified as stratification variables. Variances on parameter estimates were
estimated using Taylor Series Linearisation, the default option for PROC
SURVEYREG. For these analyses, the first plausible value of estimated student
achievement only was used. Somewhat more precise estimates of the standard errors
around parameters may have been obtained with the incorporation of the five plausible
values available in the database, however for these exploratory investigations aimed at
improving adjustments for non-response, the use of the first plausible value only was
considered sufficient.

4.4.4 Sampling Frame variables
The first model examines average student mathematics achievement for a school
explained by the following school level variables, all available in the Sampling Frame:

- The TIMSS stratification variables, State and Territory, Sector and Location
  (NSW, Government Schools and Metropolitan locations were set respectively
  as references);
- The size of the school (‘smallest’ = 25th percentile of enrolment sizes;
  ‘midsized’ = middle 50 percent; ‘largest’ = top 25th percentile);
- The school-postcode based SEIFA IEO score;
- The proportion of female students at the school;
- The proportion of ATSI students at Year 8.

Estimates for the regression coefficients for each variable and their standard errors
within this model are presented in Table 4-12. The proportion of variance in
mathematics achievement explained by this combination of variables is presented at the
bottom of the table.
Table 4-12: Multiple regression on mathematics achievement, TIMSS 2011 Year 8, Model 1: Sampling Frame variables

<table>
<thead>
<tr>
<th>Parameter Group</th>
<th>Parameter</th>
<th>Source</th>
<th>Estimate</th>
<th>StdErr</th>
<th>DenDF</th>
<th>tValue</th>
<th>Probt</th>
<th>Standardised Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>1</td>
<td>86.73</td>
<td>69.72</td>
<td>222</td>
<td>1.24</td>
<td>0.21</td>
<td>0.00</td>
</tr>
<tr>
<td>State</td>
<td>NSW (reference)</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>VIC</td>
<td>1</td>
<td>-20.37</td>
<td>12.29</td>
<td>222</td>
<td>-1.66</td>
<td>0.10</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td>QLD</td>
<td>1</td>
<td>-15.42</td>
<td>11.71</td>
<td>222</td>
<td>-1.32</td>
<td>0.19</td>
<td>-0.08</td>
</tr>
<tr>
<td></td>
<td>SA</td>
<td>1</td>
<td>-28.86</td>
<td>10.38</td>
<td>222</td>
<td>-2.78</td>
<td>0.01</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>WA</td>
<td>1</td>
<td>-32.61</td>
<td>14.14</td>
<td>222</td>
<td>-2.31</td>
<td>0.02</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td>TAS</td>
<td>1</td>
<td>-14.79</td>
<td>11.90</td>
<td>222</td>
<td>-1.24</td>
<td>0.22</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>NT</td>
<td>1</td>
<td>-38.67</td>
<td>20.32</td>
<td>222</td>
<td>-1.90</td>
<td>0.06</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>ACT</td>
<td>1</td>
<td>-38.81</td>
<td>17.53</td>
<td>222</td>
<td>-2.21</td>
<td>0.03</td>
<td>-0.07</td>
</tr>
<tr>
<td>Sector</td>
<td>Government (reference)</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Catholic</td>
<td>1</td>
<td>14.95</td>
<td>10.91</td>
<td>222</td>
<td>1.37</td>
<td>0.17</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Independent</td>
<td>1</td>
<td>34.40</td>
<td>17.65</td>
<td>222</td>
<td>1.95</td>
<td>0.05</td>
<td>0.15</td>
</tr>
<tr>
<td>Location</td>
<td>Metro (reference)</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Provincial</td>
<td>1</td>
<td>21.83</td>
<td>9.08</td>
<td>222</td>
<td>2.40</td>
<td>0.02</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td>1</td>
<td>38.91</td>
<td>32.53</td>
<td>222</td>
<td>1.20</td>
<td>0.23</td>
<td>0.05</td>
</tr>
<tr>
<td>School Type</td>
<td>Secondary (reference)</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1</td>
<td>19.82</td>
<td>10.41</td>
<td>222</td>
<td>1.90</td>
<td>0.06</td>
<td>0.10</td>
</tr>
<tr>
<td>School Size</td>
<td>Large (reference)</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>1</td>
<td>10.19</td>
<td>10.15</td>
<td>222</td>
<td>1.00</td>
<td>0.32</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>1</td>
<td>-27.22</td>
<td>11.93</td>
<td>222</td>
<td>-2.28</td>
<td>0.02</td>
<td>-0.13</td>
</tr>
<tr>
<td>Composition</td>
<td>Proportion of girls</td>
<td>1</td>
<td>13.66</td>
<td>27.87</td>
<td>222</td>
<td>0.49</td>
<td>0.62</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>propATSI</td>
<td>1</td>
<td>-1.14</td>
<td>0.87</td>
<td>222</td>
<td>-1.30</td>
<td>0.19</td>
<td>-0.07</td>
</tr>
<tr>
<td>Socio-Economic</td>
<td>SEIFA IEO</td>
<td>1</td>
<td>0.41</td>
<td>0.07</td>
<td>222</td>
<td>6.14</td>
<td>0.00</td>
<td>0.41</td>
</tr>
</tbody>
</table>

The ‘Source’ column indicates the data source. In model 1, all variables are from source ‘1’, the Sampling Frame.

This model explained approximately 26% of the variation in the mathematics outcome measure. A number of the parameters were significant. When comparing the standardised regression coefficients, the most influential variable is the school postcode SEIFA IEO measure. The Independent school sector showed higher scores on mathematics achievement compared to students from government schools after conditioning for socio-economic background, location and the other variables. The outcomes for students in the smallest schools was lower than for those from the largest
schools after controlling across this set of variables. The positive parameter estimates
for the provincial location appears spurious. This same effect was observed in Mills
and Voon’s (2011) analysis described in Section 3.3.1. These results are possibly
explained by the relatively small sample sizes and/or associations between the
explanatory variables, for example with the IEO measure capturing the resource
challenges that can be associated with location.

In order to evaluate the relative importance of the SEIFA IEO measure in explaining
achievement compared to the other sampling frame variables, a supplementary analysis
examined mathematics achievement explained by this variable alone. 18.5% of the
variation in mathematics achievement was explained by variation in the IEO of the
location of the school, around 75% of the explanatory power of model 1 was explained
by this single variable.

4.4.5 Addition of MySchool (school level) variables
The second model adds to the variables included in the first model the following
variables, all from MySchool:

- The school mean Numeracy score for the Year 9 NAPLAN assessments from
  2010. While that measure was taken of an older cohort than TIMSS (Year 8 in
  2011), it is likely that a good predictor of the mean NAPLAN numeracy
  performance of the TIMSS students is the mean score of this older cohort,
  because the students from the respective cohorts will have been drawn from the
  same family backgrounds, will have experienced the same approaches to
  schooling, the same teachers, and other influences. For some Australian
  jurisdictions it would have been possible to use the 2010 NAPLAN numeracy
  means from the Year 7 group which is the same cohort from which the TIMSS
sample was drawn in 2011, however, for Queensland, Western Australia and South Australia Year 7 was the last year of primary schooling. The school mean NAPLAN scores at Year 7 for these jurisdictions could therefore not be directly linked to the Year 8 TIMSS schools those students had moved to in the previous year.

- The Index of Community Socio-educational Advantage (ICSEA) score for the school. A discussion of this variable was provided in Chapter 3.

- The student attendance rate:

  “The student attendance rate is defined as the number of actual full-time equivalent student-days attended by full-time students in Years 1 to 10 as a percentage of the total number of possible student-days attended over the (reporting) period. The student attendance rate information is collected by schools and reported on My School twice yearly by Indigenous status for Semester 1 (Terms 1 and 2) and Term 3. (ACARA (2017c))

- The proportion of students with language backgrounds other than English (LBOTE)

- Total reported recurrent funding per student. This is the sum of four per student recurrent funding measures reported on MySchool: Australian Government recurrent funding; State/Territory government recurrent funding; Fees, charges and parent contributions; and Other private sources

Estimates for the regression coefficients for each variable and their standard errors within this model are presented in Table 4-13.
Table 4-13: Multiple regression on mathematics achievement, TIMSS 2011 Year 8, Model 2: Addition of MySchool variables

<table>
<thead>
<tr>
<th>Parameter Group</th>
<th>Parameter</th>
<th>Source</th>
<th>Estimate</th>
<th>StdErr</th>
<th>DenDF</th>
<th>tValue</th>
<th>Probt</th>
<th>Standardised Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td></td>
<td>-260.58</td>
<td>119.84</td>
<td>218</td>
<td>-2.17</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>State</td>
<td>NSW (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>VIC</td>
<td>1</td>
<td>-22.40</td>
<td>10.38</td>
<td>218</td>
<td>-2.16</td>
<td>0.03</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>QLD</td>
<td>1</td>
<td>-0.78</td>
<td>11.32</td>
<td>218</td>
<td>-0.07</td>
<td>0.94</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>SA</td>
<td>1</td>
<td>-14.37</td>
<td>10.02</td>
<td>218</td>
<td>-1.43</td>
<td>0.15</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>WA</td>
<td>1</td>
<td>-15.56</td>
<td>13.94</td>
<td>218</td>
<td>-1.12</td>
<td>0.27</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>TAS</td>
<td>1</td>
<td>-13.89</td>
<td>12.31</td>
<td>218</td>
<td>-1.13</td>
<td>0.26</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>NT</td>
<td>1</td>
<td>-35.07</td>
<td>18.07</td>
<td>218</td>
<td>-1.94</td>
<td>0.05</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td>ACT</td>
<td>1</td>
<td>-7.81</td>
<td>15.42</td>
<td>218</td>
<td>-0.51</td>
<td>0.61</td>
<td>-0.01</td>
</tr>
<tr>
<td>Sector</td>
<td>Government (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Catholic</td>
<td>1</td>
<td>7.09</td>
<td>10.38</td>
<td>218</td>
<td>0.68</td>
<td>0.50</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Independent</td>
<td>1</td>
<td>13.30</td>
<td>14.62</td>
<td>218</td>
<td>0.91</td>
<td>0.36</td>
<td>0.06</td>
</tr>
<tr>
<td>Location</td>
<td>Metro (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Provincial</td>
<td>1</td>
<td>23.76</td>
<td>9.65</td>
<td>218</td>
<td>2.46</td>
<td>0.01</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td>1</td>
<td>25.20</td>
<td>35.30</td>
<td>218</td>
<td>0.71</td>
<td>0.48</td>
<td>0.03</td>
</tr>
<tr>
<td>School Type</td>
<td>Secondary (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1</td>
<td>17.10</td>
<td>10.59</td>
<td>218</td>
<td>1.61</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>School Size</td>
<td>Large (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>1</td>
<td>12.32</td>
<td>10.16</td>
<td>218</td>
<td>1.21</td>
<td>0.23</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>1</td>
<td>-17.28</td>
<td>14.36</td>
<td>218</td>
<td>-1.20</td>
<td>0.23</td>
<td>-0.08</td>
</tr>
<tr>
<td>Composition</td>
<td>Proportion of girls</td>
<td>1</td>
<td>6.46</td>
<td>21.06</td>
<td>218</td>
<td>0.31</td>
<td>0.76</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>propATSI</td>
<td>1</td>
<td>0.67</td>
<td>0.98</td>
<td>218</td>
<td>0.69</td>
<td>0.49</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>propLBOTE</td>
<td>2</td>
<td>0.36</td>
<td>0.22</td>
<td>218</td>
<td>1.63</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Socio-Economic Factors</td>
<td>SEIFA IEO</td>
<td>1</td>
<td>0.22</td>
<td>0.08</td>
<td>218</td>
<td>2.59</td>
<td>0.01</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>School ICSEA</td>
<td>2</td>
<td>0.01</td>
<td>0.17</td>
<td>218</td>
<td>0.05</td>
<td>0.96</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Total recurrent funding per student</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>218</td>
<td>-0.21</td>
<td>0.83</td>
<td>-0.02</td>
</tr>
<tr>
<td>Achievement</td>
<td>NAPLAN Y9 Numeracy school mean</td>
<td>2</td>
<td>0.67</td>
<td>0.18</td>
<td>218</td>
<td>3.64</td>
<td>0.00</td>
<td>0.30</td>
</tr>
<tr>
<td>School Climate</td>
<td>Attendance Rate</td>
<td>2</td>
<td>1.38</td>
<td>1.55</td>
<td>218</td>
<td>0.89</td>
<td>0.37</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Most variables in the model are from the Sampling Frame (source 1); Source ‘2’ indicates the MySchool school-level variables.

This model has led to an increase in the explained variation in mathematics outcomes of an additional 6 percentage points to 32%. The explanatory variable ‘Year 9 school Numeracy mean’ has the largest standardised regression coefficient (and smallest p-
value), and the SEIFA IEO score also remained a strong predictor of achievement. The ICSEA measure was not a significant predictor. This is likely explained by its association with other variables, particularly socio-economic background as measured by the IEO score, the proportion of ATSI students, and the school geographical location, all of which are components of ICSEA.

**The Variance Inflation Factor**

To explore this association further, the Variance Inflation Factor (VIF) statistic was calculated for the model variables:

\[
VIF = \frac{1}{(1 - R^2_k)}
\]

where \(R^2_k\) is the coefficient of determination for regression of the \(i\)th explanatory variable on all of the other explanatory variables \(X_k = X_{other}\). The component \(1 - R^2_k\) is known as the Tolerance (H. Park (2003)). The VIF is a measure of how much the variance of an estimated regression coefficient has increased through its association with other variables in the regression model. When correlations between the explanatory variables are very high, it can lead to instability in the estimation of regression coefficients, known as ‘multi-collinearity’. Under multi-collinearity, estimates may for example appear implausible in magnitude or have negative sign when positive coefficients would be expected substantively. A VIF value larger than 10 is considered by some practitioners as worthy of further investigation, for example Schreiber-Gregory (2017).

The VIF statistic for many of the variables in model 2, for example State, location, school type, size and recurrent funding had values less than 2. The Year 10 Numeracy
Mean and IEO measures had VIF statistics of around 5. The school ICSEA measure in this model was close to 15. This high value suggests that there was a redundancy in the inclusion of this variable and the other variables reflecting socio-economic status. With a number of variables related to social and economic background factors available, it was important to explore models with various combinations of these variables to see which combination seemed to function best in explaining achievement. Some examples of the models explored are presented in the next section where student level variables are also added to the explanatory models.

4.4.6 Addition of student Level variables

Student level variables were added to the model explaining average student mathematics achievement for a school as described below.

**Student socio-economic background**

The addition of student level variables (‘source 3’) begins with a student level indicator of socio-economic background. The student level variables that contribute to the ICSEA score were not available for this thesis, nor were student postcodes that could be used to produce a student level SEIFA measure obtained from the area in which they live such as IEO. However, questions about parental education, and educational resources in the home were part of the TIMSS Student Questionnaire, and these were used to produce a scale of ‘Home Educational Resources’ (HER), which is used in analyses of TIMSS data as a measure of the socio-economic background of the student.

Figure 4-1 shows the items from the TIMSS student questionnaire that contribute to the Home Educational Resources scale at eighth grade (Martin and Mullis (2012)).
Valid responses were required for at least two of these items to produce a HER score for participating students. Of 7556 students in the Australian sample, 7147 provided a valid response to the question about highest level of education, 7402 responded to the item about number of books in the home and 7426 provided information about home study supports. The TIMSS Methods and Procedures resource Martin and Mullis (2012) provides further details of the construction of the HER scale. A HER score was produced for 7402 students overall.

In the absence of student level SEIFA or ICSEA based measures HER will be used as a proxy for a student level background measure potentially accessible as data that could contribute to better non-response management and stronger analyses of large-scale survey data.

Model 3 in Table 4-14 shows the effect of inclusion of the HER score in the model to explain mathematics achievement. HER is a statistically significant factor and the standardised regression coefficients indicate that it is contributing also equally to the school NAPLAN mean, the other main contributor to the model. Also added is student sex, and this also is a statistically significant contributor to explaining achievement, with girls estimated as scoring on average around 8 points lower than boys, all other
variables equal. With the inclusion of the student level HER measure, the school level measure of socio-economic advantage, ICSEA was not statistically significant. Any significant differences at the State and sector level observed in the first model have disappeared with the inclusion of these additional variables.

The percentage of variance explained with this model was 40%, an 8 percentage point increase on the previous model.
As observed in the previous analysis there were a number of different socio-economic background variables available – the school postcode based SEIFA IEO, the school level ICSEA score, the student level HER variable, and other related factors such as the proportion of Indigenous students - and there was evidence of possible multicollinearity.
occurring between the variables. Various combinations of these variables were explored under different models to determine the best combination of variables to use.

The following table shows an extract for each of the combination of variables explored in models labelled 3a to 3e. The other background variables such as State and sector discussed in the models above, and also the NAPLAN mean Numeracy score as a prior performance measure are not shown in this table, although they also were incorporated into the models.

**Table 4-15: Multiple regression on mathematics achievement (extract), TIMSS 2011 Year 8, Models 3a-3e: socio-economic variables**

<table>
<thead>
<tr>
<th>Parameter Group</th>
<th>Parameter</th>
<th>Source</th>
<th>Models -</th>
<th>3</th>
<th>3a</th>
<th>3b</th>
<th>3c</th>
<th>3d</th>
<th>3e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Metro (reference)</td>
<td>1</td>
<td></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Provincial</td>
<td>1</td>
<td></td>
<td></td>
<td>0.11</td>
<td>x</td>
<td>0.10</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td>1</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Composition</td>
<td>propATSI</td>
<td>1</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>propLBOTE</td>
<td>2</td>
<td></td>
<td></td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Socio-Economic Factors</td>
<td>SEIFA IEO</td>
<td>2</td>
<td></td>
<td></td>
<td>0.19</td>
<td></td>
<td></td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>School ICSEA</td>
<td>2</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Home Educational Re:</td>
<td>3</td>
<td></td>
<td></td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Total recurrent fundin</td>
<td>2</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models -</th>
<th>R-sqd</th>
<th>Adj R-sqd</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>3a</td>
<td>0.36</td>
<td>0.38</td>
</tr>
<tr>
<td>3b</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>3c</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>3d</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>3e</td>
<td>0.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

In columns 3a to 3e, if the cell is blank, the variable was not included in the model. If the cell has a cross, the variable was in the model, but the estimated regression coefficient was not statistically significantly different from zero. If the variable was in the model and significant, then the standardised regression coefficient is shown. For example, model 3b involves both HER which is a significant factor with a standardised regression coefficient of 0.27, and ICSEA which was not significant in this model. The
IEO score is not included, nor the location variable, the proportion of ATSI and the proportion of LBOTE. (These variables form part of ICSEA). Total recurrent funding is also not included in the model. The periods in the ‘Metro’ row indicate that metro was the reference category for the location classification variable (which appeared in all models above except 3b).

One clear observation is that all of the models explored have similar R-squared values, indicating that all combinations have a similar degree of explanatory power. When ICSEA is included as the only socio-economic factor (model 3e) it is a significant contributor to the model. In the other models HER and/or IEO appear to be performing an equivalent job in explaining achievement. Because of the desire to have a student level socio-economic variable for use later in the thesis, it was decided to prefer the models involving the HER measure and/or IEO, and not persist with the (school level) ICSEA variable.

**Variance inflation statistics by model**

Table 17 shows the VIF statistics calculated for the variables reflecting socio-economic background for these models. The high VIFs observed in the earlier models when both IEO and ICSEA are included in the models are brought down well below the guideline of ‘10’ (page 148) when one of these is removed. This provides further support for the decision to remove ICSEA from later models.

**Table 4-16: Variance Inflation Statistics on socio-economic variables by regression model**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>2</th>
<th>3</th>
<th>3a</th>
<th>3b</th>
<th>3c</th>
<th>3d</th>
<th>3e</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEIFA IEO</td>
<td>4.86</td>
<td>4.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.22</td>
</tr>
<tr>
<td>School ICSEA</td>
<td>14.84</td>
<td>14.99</td>
<td>6.88</td>
<td>5.67</td>
<td></td>
<td></td>
<td>6.65</td>
</tr>
<tr>
<td>Home Educational Resources</td>
<td>1.21</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.20</td>
<td>1.17</td>
</tr>
<tr>
<td>Total recurrent funding per student</td>
<td>1.78</td>
<td>1.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Performance on science as a proxy for prior performance

Model 4 in Table 4-17 includes the addition of the student science outcome as a predictor for mathematics achievement. As explained earlier, this variable is being treated as a proxy for a recent NAPLAN performance outcome in Numeracy for the student, an indicator of mathematics ability that would likely be highly correlated with the expected performance on the TIMSS mathematics assessment and could be used in non-response adjustment. The model retains the student-level HER measure, but removes the school level ICSEA
The proportion of variance explained by this model rises substantially to 74%. The correlation between the mathematics and science outcomes contributes largely to the explanatory power of the model. Some other variables that are as statistically significant in this model include the school Numeracy mean, a contextual effect, and the proportion of students with language backgrounds other than English.

Two other notable observations from this model are:
• The regression coefficient for the socio-economic background variable HER is not significantly different from zero, suggesting that the impact of socio-economic background is already manifested in prior achievement;
• The school mean Numeracy score remains as a significant predictor, even when the student level performance measure is added to the model. This contextual effect indicates the potential of the school environment to add to student achievement outcomes.

Variance inflation statistics were all below 4, indicating no multicollinearity concerns.

To investigate these outcomes further, a new model was explored that removed the mean numeracy score for the school. The outcomes of this model are presented in Table 4-18. This model explains almost as much of the variation in achievement as the model above (73% instead of 74%). The HER measure is now a significant factor. The attendance rate is also a significant factor and has a positive regression coefficient, as expected. These outcomes and those made in relation to Model 4 may be tapping into the observations of Lamb et al. (2004) that it is not just a demonstrated prior ability in maths that predicts higher outcomes, but the school climate effects of having a cohort of students with a stronger performance background.
Table 4-18: Multiple regression on mathematics achievement, TIMSS 2011 Year 8, Model 4a: removal of school NAPLAN Numeracy mean

<table>
<thead>
<tr>
<th>Parameter Group</th>
<th>Parameter</th>
<th>Source</th>
<th>Estimate</th>
<th>StdErr</th>
<th>DenDF</th>
<th>tValue</th>
<th>Probt</th>
<th>Standardised Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>-75.61</td>
<td>40.72</td>
<td>234</td>
<td>-1.86</td>
<td>0.06</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>State</td>
<td>NSW (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>VIC</td>
<td>1</td>
<td>-0.21</td>
<td>4.02</td>
<td>234</td>
<td>-0.05</td>
<td>0.96</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>QLD</td>
<td>1</td>
<td>4.06</td>
<td>4.23</td>
<td>234</td>
<td>0.96</td>
<td>0.34</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>SA</td>
<td>1</td>
<td>-2.19</td>
<td>3.71</td>
<td>234</td>
<td>-0.59</td>
<td>0.56</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>WA</td>
<td>1</td>
<td>-4.05</td>
<td>4.46</td>
<td>234</td>
<td>-0.91</td>
<td>0.36</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>TAS</td>
<td>1</td>
<td>-1.62</td>
<td>3.91</td>
<td>234</td>
<td>-0.41</td>
<td>0.68</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>NT</td>
<td>1</td>
<td>-1.58</td>
<td>6.71</td>
<td>234</td>
<td>-0.24</td>
<td>0.81</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ACT</td>
<td>1</td>
<td>3.24</td>
<td>4.42</td>
<td>234</td>
<td>0.73</td>
<td>0.46</td>
<td>0.01</td>
</tr>
<tr>
<td>Sector</td>
<td>Government (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Catholic</td>
<td>1</td>
<td>2.31</td>
<td>3.72</td>
<td>234</td>
<td>0.62</td>
<td>0.54</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Independent</td>
<td>1</td>
<td>10.55</td>
<td>6.61</td>
<td>234</td>
<td>1.60</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>School Type</td>
<td>Secondary (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1</td>
<td>2.48</td>
<td>4.32</td>
<td>234</td>
<td>0.58</td>
<td>0.57</td>
<td>0.01</td>
</tr>
<tr>
<td>School Size</td>
<td>Large (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>1</td>
<td>2.17</td>
<td>3.61</td>
<td>234</td>
<td>0.60</td>
<td>0.55</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>1</td>
<td>-5.97</td>
<td>4.51</td>
<td>234</td>
<td>-1.32</td>
<td>0.19</td>
<td>-0.03</td>
</tr>
<tr>
<td>Composition</td>
<td>Proportion of girls</td>
<td>1</td>
<td>4.67</td>
<td>8.49</td>
<td>234</td>
<td>0.55</td>
<td>0.58</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>propLBOTE</td>
<td>2</td>
<td>0.35</td>
<td>0.08</td>
<td>234</td>
<td>4.48</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>Socio-Economic</td>
<td>Factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Home Educational Resources</td>
<td>3</td>
<td>1.43</td>
<td>0.53</td>
<td>234</td>
<td>2.69</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Sex</td>
<td>Girls</td>
<td>3</td>
<td>3.47</td>
<td>1.78</td>
<td>234</td>
<td>1.95</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Boys (reference)</td>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Student achievement in Science</td>
<td>3</td>
<td>0.79</td>
<td>0.01</td>
<td>234</td>
<td>54.48</td>
<td>0.00</td>
<td>0.80</td>
</tr>
<tr>
<td>School Climate</td>
<td>Attendance Rate</td>
<td>2</td>
<td>1.49</td>
<td>0.45</td>
<td>234</td>
<td>3.30</td>
<td>0.00</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Additional school climate factors

To finalise the investigation of the factors relating to outcomes on the TIMSS mathematics assessment, the following model draws additional information about student attitudes and the school climate, drawn from the background data collected through School, Student and Teacher questionnaires that form part of the instrumentation for TIMSS (see section 4.2.1).
Sixteen different scales are included in the TIMSS international database which are drawn from items from these instruments. Those scales are presented in Table 4-19, along with frequencies and descriptive statistics based on the Australian Year 8 data.

Table 4-19: Developed scales based on student, school and teacher questionnaires: TIMSS 2011

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>Label</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>*HOME EDUCATIONAL RESOURCES</td>
<td>7402</td>
<td>11.16</td>
<td>1.63</td>
<td>4.32</td>
<td>14.02</td>
</tr>
<tr>
<td>Student</td>
<td>*STUDENTS BULLIED AT SCHOOL</td>
<td>7377</td>
<td>9.88</td>
<td>1.93</td>
<td>2.90</td>
<td>13.04</td>
</tr>
<tr>
<td>Student</td>
<td>*STUDENTS LIKE LEARNING MATHEMATICS</td>
<td>7389</td>
<td>9.29</td>
<td>1.69</td>
<td>5.04</td>
<td>13.47</td>
</tr>
<tr>
<td>Student</td>
<td>*STUDENTS VALUE LEARNING MATHEMATICS</td>
<td>7374</td>
<td>9.96</td>
<td>1.83</td>
<td>3.41</td>
<td>13.71</td>
</tr>
<tr>
<td>Student</td>
<td>*STUDENT CONFIDENCE WITH MATHEMATICS</td>
<td>7372</td>
<td>10.15</td>
<td>2.05</td>
<td>3.18</td>
<td>15.82</td>
</tr>
<tr>
<td>Student</td>
<td>*STUDENTS ENGAGED IN MATHEMATICS LESSONS</td>
<td>7374</td>
<td>9.24</td>
<td>1.71</td>
<td>3.27</td>
<td>14.34</td>
</tr>
<tr>
<td>School</td>
<td>*INSTRUCTION AFFECTED BY MATHEMATICS RESOURCE SHORTAGES</td>
<td>6841</td>
<td>11.11</td>
<td>1.88</td>
<td>3.09</td>
<td>15.23</td>
</tr>
<tr>
<td>School</td>
<td>*SCHOOL EMPHASIS ON ACADEMIC SUCCESS - PRINCIPAL REPORTS</td>
<td>6788</td>
<td>10.72</td>
<td>2.37</td>
<td>4.91</td>
<td>15.57</td>
</tr>
<tr>
<td>School</td>
<td>*SCHOOL DISCIPLINE AND SAFETY</td>
<td>6799</td>
<td>9.93</td>
<td>1.46</td>
<td>6.01</td>
<td>13.94</td>
</tr>
<tr>
<td>Teacher</td>
<td>*TEACHERS REPORT PROBLEMS WITH WORKING CONDITIONS</td>
<td>4716</td>
<td>10.71</td>
<td>2.08</td>
<td>3.10</td>
<td>14.89</td>
</tr>
<tr>
<td>Teacher</td>
<td>*SCHOOL EMPHASIS ON ACADEMIC SUCCESS - TEACHER REPORTS</td>
<td>4729</td>
<td>10.33</td>
<td>2.43</td>
<td>4.99</td>
<td>16.21</td>
</tr>
<tr>
<td>Teacher</td>
<td>*SAFE AND ORDERLY SCHOOL</td>
<td>4708</td>
<td>10.41</td>
<td>2.25</td>
<td>4.58</td>
<td>13.22</td>
</tr>
<tr>
<td>Teacher</td>
<td>*CONFIDENCE IN TEACHING MATHEMATICS</td>
<td>4584</td>
<td>9.98</td>
<td>2.05</td>
<td>5.07</td>
<td>11.99</td>
</tr>
<tr>
<td>Teacher</td>
<td>*TEACHER CAREER SATISFACTION</td>
<td>4700</td>
<td>9.73</td>
<td>2.22</td>
<td>1.82</td>
<td>13.80</td>
</tr>
<tr>
<td>Teacher</td>
<td>*COLLABORATE TO IMPROVE TEACHING</td>
<td>4705</td>
<td>10.32</td>
<td>2.03</td>
<td>4.85</td>
<td>15.77</td>
</tr>
<tr>
<td>Teacher</td>
<td>*INSTRUCTION TO ENGAGE STUDENTS</td>
<td>4644</td>
<td>9.88</td>
<td>1.87</td>
<td>2.32</td>
<td>11.54</td>
</tr>
</tbody>
</table>

As the background data for schools and teachers can be linked to individual student records, all scales can potentially be used as explanatory variables in explaining mathematics achievement.
As can be observed from the table, whereas the Student Questionnaire scales appear for all, or almost all students in the database, there is a considerable amount of missing data, particularly in relation to the Teacher Questionnaire scales. The Teacher Questionnaire scales showed minimal power for explaining achievement. A number of the above scales appeared to contribute to the explanatory power of the model. The variables with the most explanatory power - school emphasis on academic success; school discipline and safety; like learning mathematics; confidence with mathematics; and engaged in mathematics lessons – have been added to regression model 5, as shown in Table 4-20.
Table 4-20: Multiple regression on mathematics achievement, TIMSS 2011 Year 8, Model 5: addition of school climate and student affect variables

<table>
<thead>
<tr>
<th>Parameter Group</th>
<th>Parameter</th>
<th>Source</th>
<th>Estimate</th>
<th>StdErr</th>
<th>DenDF</th>
<th>tValue</th>
<th>Probt</th>
<th>Standardised Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td></td>
<td>-138.99</td>
<td>28.36</td>
<td>207</td>
<td>-4.90</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>State</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NSW (reference)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VIC</td>
<td>1</td>
<td>-2.37</td>
<td>3.28</td>
<td>207</td>
<td>-0.72</td>
<td>0.47</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>QLD</td>
<td>1</td>
<td>-1.79</td>
<td>3.31</td>
<td>207</td>
<td>-0.54</td>
<td>0.59</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>SA</td>
<td>1</td>
<td>-5.48</td>
<td>2.79</td>
<td>207</td>
<td>-1.96</td>
<td>0.05</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>WA</td>
<td>1</td>
<td>-6.04</td>
<td>3.23</td>
<td>207</td>
<td>-1.87</td>
<td>0.06</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>TAS</td>
<td>1</td>
<td>-7.47</td>
<td>2.68</td>
<td>207</td>
<td>-2.79</td>
<td>0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>NT</td>
<td>1</td>
<td>-2.61</td>
<td>4.72</td>
<td>207</td>
<td>-0.55</td>
<td>0.58</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>ACT</td>
<td>1</td>
<td>0.82</td>
<td>4.22</td>
<td>207</td>
<td>0.20</td>
<td>0.85</td>
<td>0.00</td>
</tr>
<tr>
<td>Sector</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Government (reference)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Catholic</td>
<td>1</td>
<td>-0.95</td>
<td>2.95</td>
<td>207</td>
<td>-0.32</td>
<td>0.75</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Independent</td>
<td>1</td>
<td>5.58</td>
<td>5.63</td>
<td>207</td>
<td>0.99</td>
<td>0.32</td>
<td>0.03</td>
</tr>
<tr>
<td>School Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Secondary (reference)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1</td>
<td>1.18</td>
<td>3.34</td>
<td>207</td>
<td>0.35</td>
<td>0.73</td>
<td>0.01</td>
</tr>
<tr>
<td>School Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Large (reference)</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>1</td>
<td>3.65</td>
<td>2.92</td>
<td>207</td>
<td>1.25</td>
<td>0.21</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>1</td>
<td>-4.94</td>
<td>3.99</td>
<td>207</td>
<td>-1.24</td>
<td>0.22</td>
<td>-0.03</td>
</tr>
<tr>
<td>Composition</td>
<td>Proportion of girls</td>
<td>1</td>
<td>4.02</td>
<td>7.84</td>
<td>207</td>
<td>0.51</td>
<td>0.61</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>propLBOTE</td>
<td>2</td>
<td>0.21</td>
<td>0.07</td>
<td>207</td>
<td>3.02</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Socio-Economic Factors</td>
<td>Home Educational Resources</td>
<td>3</td>
<td>0.65</td>
<td>0.35</td>
<td>207</td>
<td>1.86</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Girls</td>
<td>3</td>
<td>8.22</td>
<td>1.39</td>
<td>207</td>
<td>5.91</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Boys (reference)</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achievement</td>
<td>Student achievement in Science</td>
<td>3</td>
<td>0.70</td>
<td>0.01</td>
<td>207</td>
<td>51.50</td>
<td>0.00</td>
<td>0.70</td>
</tr>
<tr>
<td>School Climate</td>
<td>Attendance Rate</td>
<td>2</td>
<td>1.54</td>
<td>0.26</td>
<td>207</td>
<td>5.95</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>School emphasis on academic success (SCH Q)</td>
<td>2</td>
<td>2.41</td>
<td>0.53</td>
<td>207</td>
<td>4.56</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>School discipline and safety (SCH Q)</td>
<td>2</td>
<td>-0.26</td>
<td>1.23</td>
<td>207</td>
<td>-0.21</td>
<td>0.83</td>
<td>0.00</td>
</tr>
<tr>
<td>Student Affect</td>
<td>Like learning maths (STU Q)</td>
<td>2</td>
<td>3.55</td>
<td>0.47</td>
<td>207</td>
<td>7.50</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Confidence with maths (STU Q)</td>
<td>2</td>
<td>7.59</td>
<td>0.29</td>
<td>207</td>
<td>26.58</td>
<td>0.00</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Engaged in maths lessons (STU Q)</td>
<td>2</td>
<td>-1.22</td>
<td>0.50</td>
<td>207</td>
<td>-2.45</td>
<td>0.02</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

Along with the variables already observed as significant in explaining achievement – prior performance, socio-economic background, the proportion of students with language backgrounds other than English, and the attendance rate, the emphasis on academic success, and students’ enjoyment with, and confidence in mathematics were also statistically significant factors.
A further improvement was observed in the explanatory power of the model, up now to 78% of the variation in mathematics outcomes explained compared to 74% observed in Model 4 (Table 4-20). The incorporation of the school climate factors produces a considerable improvement on the explanatory power of the model, but unlike for HER and Science performance, each of which can readily serve as proxies for variables collected in NAPLAN or reported in MySchool to contribute to the non-response management investigations in this thesis, there are no clear variables collected through these census activities for which the TIMSS school climate factors would serve as proxies. While useful and important to be aware that, all other things equal, school climate and student affect variables can also add to the explanation of achievement in mathematics, without such variables available in a resource such as NAPLAN or MySchool, their usefulness with respect to non-response management is likely limited.

4.4.7 Summary of multiple regression analyses of mathematics achievement at the school level

Table 4-21 summarises the key outcomes of the multiple regression models investigated in this chapter.
Table 4-21: Multiple regression on mathematics achievement, TIMSS 2011 Year 8, model summary

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter</th>
<th>Source</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3a</th>
<th>3b</th>
<th>3c</th>
<th>3d</th>
<th>3e</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td></td>
<td>x</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>State</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSW</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIC</td>
<td></td>
<td>1</td>
<td>x</td>
<td>-0.11</td>
<td>-0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QLD</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td></td>
<td>1</td>
<td>-0.09</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.02</td>
</tr>
<tr>
<td>WA</td>
<td></td>
<td>1</td>
<td>-0.10</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAS</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td></td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>-0.04</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>-0.05</td>
</tr>
<tr>
<td>ACT</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Sector</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catholic</td>
<td></td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Independent</td>
<td></td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metro</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provincial</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>0.12</td>
<td>0.13</td>
<td>0.11</td>
<td>x</td>
<td>0.10</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Remote</td>
<td></td>
<td>1</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>School Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td></td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>0.01</td>
</tr>
<tr>
<td>School Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td></td>
<td>1</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle</td>
<td></td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Small</td>
<td></td>
<td>1</td>
<td>-0.13</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Composition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PropGirls</td>
<td></td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>propATSI</td>
<td></td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>propLBOTE</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socio-Economic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEIFA IEO</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>0.41</td>
<td>0.22</td>
<td>0.19</td>
<td></td>
<td>0.17</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>School ICSEA</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HER</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.28</td>
</tr>
<tr>
<td>Recurrent funding</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girls</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.05</td>
<td></td>
</tr>
<tr>
<td>Boys</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.05</td>
<td></td>
</tr>
<tr>
<td>Achievement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAPLAN Numeracy</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.30</td>
<td>0.28</td>
<td>0.24</td>
<td>0.33</td>
<td>0.26</td>
<td>0.36</td>
<td>0.25</td>
</tr>
<tr>
<td>Science Achievement</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attendance Rate</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Emphasis on success</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discipline and safety</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.05</td>
</tr>
<tr>
<td>Affect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Like maths</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>Confidence maths</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>Engaged in maths</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.03</td>
</tr>
</tbody>
</table>

| R-sqd | 0.26 | 0.32 | 0.38 | 0.37 | 0.36 | 0.38 | 0.37 | 0.37 | 0.74 | 0.79 |
| Adj R-sqd | 0.26 | 0.32 | 0.38 | 0.37 | 0.36 | 0.38 | 0.37 | 0.37 | 0.74 | 0.79 |

**table models legend**

- **x** included in the model but not significant
- `<value>` standardised regression coefficient (for a significant factor)
- `<blank>` not included in the model
- -. a reference category
If the cell is blank, the variable was not included in the model. If the cell has a cross, the variable was in the model, but the estimated regression coefficient was not statistically significantly different from zero. If the variable was in the model and statistically significant, then the standardised regression coefficient is shown.

Two strong observations from the above analyses are that:

- The set of variables currently being used for non-response management – mostly those variables available on the sampling frame – are not very powerful in explaining achievement in the context of the TIMSS survey.
- When these are supplemented with variables collected from other sources relating to socio-economic background, prior performance and school climate, the power to explain variation in achievement increases substantially. This includes variables available on MySchool, and also variables at the student level, particularly a measure of prior performance.

As was discussed in a number of sections of Chapter 2, survey variables that are related to the outcome measures of the survey are very useful in the management of non-response. In section 2.4.1, for example, there was a discussion of the need for the researcher to look at prior studies to do a qualitative evaluation of the likely differences between respondents and non-respondents on the survey outcome measures. In section 2.5.1, the degree of survey bias was quantified as the product of non-response rate and the differences between respondents and non-respondents on the outcome variable. Having variables in the survey related to outcomes will serve to evaluate these differences at the survey planning stage. At the estimation stage, making use of variables related to outcomes will clearly be valuable for the formation of weighting
classes or post-stratification (section 2.7.2); response homogeneity groups (section 2.6.3) or response propensity classes (section 2.10). The quality of the predictions used in regression models (section 2.6.6), or the use of school substitution (section 2.6.5) also rely on variables correlated to the outcome variables of the survey.

The investigation in this chapter of the TIMSS survey variables and how they relate to the outcome of mathematics achievement has identified a number of key variables that are related to mathematics outcomes, and given an indication of their relative power in explaining achievement in mathematics.

While the student level variables most likely to be useful in non-response management, for example student performances on NAPLAN and socio-economic background data linked to student location and parental education and occupation, were not available for this research, it is clear from the proxies used for these variables that such data would contribute significantly to explaining student achievement if they were made available.

Another finding from these analyses is that while the most important factors for explaining achievement are related to factors such as socio-economic background, prior performance and school climate, the interplay of particular variables across these broader themes is complex. For example, the extent to which it is the climate of the school that leads to improved performance, or whether a higher performing cohort leads to better climate – the themes explored in the papers discussed earlier by Miller and Voon (2011), Lamb et al. (2004) and Marks(2010) - is clearly very complex. While a substantive analysis of exactly how those variables interact with each other is beyond the scope of this thesis, it is clear from the analyses above that a full investigation of the
factors relating to achievement depends on deeper information about schools and their students than the data that has typically been made available for the management of non-response.

Another desirable quality of variables for non-response management is that they help to explain response. Chapter 5 investigates the TIMSS data with respect to this other desirable property of variables.
Chapter 5  Patterns of response

The previous chapter examined factors related to performance on the TIMSS mathematics assessment. If available, a measure of prior academic performance was a very important contributor to explaining variation in mathematics achievement. Student and school level socio economic factors were also found to be important. Other factors such as the proportion of Indigenous students at the school also contribute to the explanatory power. The findings of chapter 4 essentially confirmed the outcomes of the previous studies discussed in Chapter 2 (Miller and Voon (2011), Lamb et al. (2004), and Marks (2010)) with respect to the major factors that explain academic performance outcomes for Australian students.

In making decisions about the management of non-response, an important additional question is whether the non-response that has occurred for the survey is missing at random, i.e. random conditional on available variables such as those explored in Chapter 4, or whether the missingness is related to the outcome measure in ways not fully explained by those variables, i.e. ‘NMAR’, (see section 2.5.3). In the former case, there are methods available to use the variables associated with non-response to adjust for it as described in Chapter 2. For example one approach is the formation of weighting classes incorporating those important variables, within which non-response can be treated as missing completely at random (MCAR). In this case the bias will be removed and the effects of missingness on estimates will be limited to an additional component of variance associated with having a smaller number of participants in the weighting
class than was intended. Appropriate weighting will minimise the effects of non-
response bias. This is the ‘quasi-randomisation’ approach discussed in Section 2.6.2.
However when data are NMAR, even after conditioning on survey design variables,
then further modelling may be required, and the management of non-response becomes
more challenging. Some bias will remain even after non-response adjustments that have
assumed MAR.

This chapter will examine participation rates in the TIMSS survey across the variables
identified as the most important variables in explaining outcomes, to assess whether
participation appears related to outcomes in ways that are not completely addressed by
other variables from the survey.

5.1 Identifying absent students

As discussed in Chapter 3, TIMSS is one of the surveys conducted under Australia’s
National Assessment Program (NAP). Participation for surveys conducted under this
program is effectively mandatory at the school level. Table 5-1 shows that the school
response rate was in the high 90%’s, even before the use of replacement schools. Within
those schools, 90% of eligible students from the sampled classrooms participated.

<table>
<thead>
<tr>
<th>School Participation</th>
<th>Class Participation</th>
<th>Student Participation</th>
<th>Overall Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Replacement</td>
<td>After Replacement</td>
<td>Before Replacement</td>
<td>After Replacement</td>
</tr>
<tr>
<td>96%</td>
<td>98%</td>
<td>100%</td>
<td>87%</td>
</tr>
<tr>
<td>90%</td>
<td>87%</td>
<td>88%</td>
<td></td>
</tr>
</tbody>
</table>

In Chapter 4 the weighting process applied to the TIMSS survey data was outlined. A
component of the weighting was the non-response adjustment applied to participating
students within sampled classes (section 4.3.3 page 133).
The student non-participation adjustment for students sampled from class $j$ of school $i$ was calculated as the number of eligible students from that class divided by the number who participated.

$$A_{i,j}^{st} = \frac{s_{i,j}^{rs} + s_{i,j}^{nr}}{s_{i,j}^{rs}}$$

$s_{i,j}^{rs}$ is the number of participating students in the $j^{th}$ class of the $i^{th}$ school, and $s_{i,j}^{nr}$ is the number of eligible non-participating students.

This non-response adjustment is provided as a separate variable in the international database. Using this variable, the number of eligible students from each class can be calculated as a school level variable equal to the product of the number who participated and the non-participation adjustment. From this, the level of non-response could be calculated for each school.

**Additional sample of Indigenous students**

In the sample design discussion in Chapter 4, it was noted that one or two classes were sampled from each school depending on the jurisdiction, and in addition, any Indigenous students at the target grade not in those sampled classes were included in the survey. For operational reasons, these additional students were combined into a separate pseudo-class and, as they were representing themselves only, were given a student weight (prior to non-response adjustment) of 1.
Unlike the main sample of Year 8 students, the student non-response adjustment for these additional Indigenous students appears to have been applied at the stratum level, with all of these students across multiple schools in the stratum having the same adjustment weight. This approach was likely to have been used for the same reason that a similar approach was applied to the class level non-response adjustment, as explained section 4.3.2, as an attempt to avoid amplifying any bias arising from non-response at a particular school.

Because the student non-response adjustment for the additional Indigenous students was applied at the State/Territory level, it was not possible to directly translate this into the number of additional Indigenous students not responding as a school level variable, and similarly across sub-population domains (for example sector or location). These students were therefore removed from the analyses described below.

To account for these students no longer in the database, the total weight of students from the school was adjusted upwards so that the sum of the student weights from the school equalled the weighted sum of all of the students participating from the school (including the additional Indigenous students) using the weight supplied in the TIMSS database (totwgt).

5.2 Defining subpopulation categories

For the following variables, subpopulations were derived from the Sampling Frame data: State/Territory; sector (government, Catholic, independent); sex composition of the school (all girls/ all boys/coeducational); location (metropolitan /provincial/remote); SEIFA IEO quintile. In the case of the IEO quintile, these were those published by the
Australian Bureau of Statistics, categorising areas of Australia into ten equal groups. If the school was located in one of the lowest scoring 20% of areas across Australia it was in the first quintile of IEO and so on.

For a range of school-level variables, distributions were estimated from the TIMSS surveys. For each variable, the 10th, 25th, 50th, 75th and 90th percentiles were calculated. Each classification indicated the estimated proportion of the population of Year 8 students in Australia at that percentile or lower, with a further category of ‘90plus’ representing the top 10% of students in the population for each variable. The school-level variables used in this way were: ICSEA scores; NAPLAN mean Numeracy score at Year 9; School Size; Proportion of ATSI students; Proportion of LBOTE students; Attendance rate. The following student level variables were classified in the same way: TIMSS Science performance, TIMSS Home Educational Resources score.

5.3 Calculating participation rates

Weighted student participation rates were calculated for each of the subpopulation categories, thus providing population estimates of participation rates. The rates were calculated using expressions adapted from the TIMSS sampling documentation (M. F. Joncas, Pierre, 2011). The overall weighted student participation rate is the product of weighted participation rates at the school, class and student levels:

$$R_{wt}^{ovr} = R_{wt}^{sc} \cdot R_{wt}^{el} \cdot R_{wt}^{st}$$  5-2

$R_{wt}^{ovr}$ is the weighted overall participation rate, calculated over all participating schools including replacement schools.
$R_{wtd}^{sc-r}$ is the weighted school participation rate component, calculated over all students from all participating schools including replacement schools:

$$R_{wtd}^{sc-r} = \frac{\sum_{i,j}^{s+r1+r2} BW_i^{sc} \cdot FW_{i,j}^{cl} \cdot FW_{i,j}^{st}}{\sum_{i,j}^{s+r1+r2} FW_i^{sc} \cdot FW_{i,j}^{cl} \cdot FW_{i,j}^{st}}$$  

5-3

The components of $R_{wtd}^{sc-r}$ are the base weights and final weights for each of school, class and student as described in chapter 4 section 4.3.

The weighted class participation rate component of expression 5-2, $R_{wtd}^{cl}$, is calculated over all students from participating classes with at least 50 percent participation rate of students from the class, over all participating schools:

$$R_{wtd}^{cl} = \frac{\sum_{i,j}^{s+r1+r2} BW_i^{sc} \cdot BW_{i,j}^{cl} \cdot FW_{i,j}^{st}}{\sum_{i,j}^{s+r1+r2} BW_i^{sc} \cdot BW_{i,j}^{cl} \cdot FW_{i,j}^{st}}$$  

5-4

The components of $R_{wtd}^{cl}$ are the base weights and final weights for each of school, class and student as described in chapter 4 section 4.3.

The weighted student participation rate component of expression 5-2, $R_{wtd}^{st}$, is calculated over all students from participating classes with at least 50 percent participation rate of students from the class, over all participating schools:

$$R_{wtd}^{st} = \frac{\sum_{i,j}^{s+r1+r2} BW_i^{sc} \cdot BW_{i,j}^{cl} \cdot BW_{i,j}^{st}}{\sum_{i,j}^{s+r1+r2} BW_i^{sc} \cdot BW_{i,j}^{cl} \cdot FW_{i,j}^{st}}$$  

5-5

As previously, the individual components of this expression are the base weights and final weights described in section 4.3.
5.4 Participation rates across population sub-domains

The following charts show variations in response rates across the estimated student distributions across the range of variables discussed above, i.e. socio-economic background, prior performance, school composition factors, school size, attendance rates, State, sector, location, and school sex composition.

In the first instance, the variation in the participation rates against the variables are intended to provide a brief check of the key influences on non-response. Should any such variables be uncovered, that have not already been identified as potentially important with respect to non-response management, tests of statistical significance would be used to check that apparent differences between subgroups of the variable are significantly different in their response pattern.

Note that the participation rates are a product of both school- and student-level response, but because of the very high rates of school response (section 4.2.2), they mostly reflect student-level participation. Note also that they reflect patterns of participation for students sampled for TIMSS across Australia as a whole. The investigations of non-response approaches in the following chapters will be focussed on the subpopulation of students being educated in Victorian government schools.

While the response patterns across variables shown in this section are sufficient to motivate the simulations used in later chapters of this thesis, an interesting analysis for
further research would be to examine the interactions between the variables through a regression model.

5.4.1 Socio-economic background measures

Figure 5-1 shows participation rates by school-level ICSEA scores and by percentile groupings of the student-level HER measure.

![Participation rates and socio-economic background measures](image)

Figure 5-1: Participation rates by ICSEA and HER, TIMSS 2011 Year 8, Australia

Figure 5-2 shows participation rates by quintiles of the school-level IEO measure.
A general upward trend in response is apparent across the three socio-economic background variables under consideration for this project – the school-postcode based SEIFA IEO index; the ICSEA measure of socio-educational advantage published on MySchool and the student level measure of Home Educational Resources from the TIMSS survey. Response rates for students in the lowest decile for ICSEA for example are more than 10% lower than those in the highest decile.

5.4.2 Performance measures
Figure 5-3 presents participation rates against two prior performance measures: the school-level NAPLAN Numeracy mean and the student-level science score of students who participated in TIMSS.
Once again, a pattern of increasing participation is observed with higher levels of performance, as indicated by the student distribution based on the school level NAPLAN Numeracy mean, and on the estimated population distribution of achievement based on the science scores of students participating in the TIMSS assessment.

5.4.3 Participation against performance crossed with socio-economic profile

Figure 5-4 shows participation rates across groups defined as the cross product of the categories of the NAPLAN Numeracy mean with the SEIFA IEO quintiles. The label
‘p10_1’ means between the 0-10th percentile on the NAPLAN Numeracy mean and in the first SEIFA IEO quintile, and so on).

While a clear overall upward trend with increasing categories of performance, participation rates vary for SEIFA IEO quintiles within those categories of performance. The chart indicates a complex interaction of participation rates between these two variables that each contributed to explaining achievement.
5.4.4 School factors (1) – proportion of ATSI and LBOTE students

Figure 5-5 shows participation rates across two components of school composition, the proportion of ATSI students and the proportion students with language backgrounds other than English (LBOTE).

![Participation rates and proportion of ATSI; proportion of LBOTE](image)

Figure 5-5: Participation rates by ATSI, LBOTE proportions, TIMSS 2011 Year 8, Australia

A slight upward trend in participation is evident in participation rates as the proportion LBOTE students increases. In contrast, the participation rate declines as the proportion of ATSI students increases.

5.4.5 School factors (2) – school size; attendance rate

Figure 5-6 shows participation rates by school size, and by attendance rate.
There is small variation in participation rates across different school sizes. Rates of participation in TIMSS increase for schools with increasing overall attendance rates.

5.4.6 School factors (3) – school sex composition, school location

Figure 5-7 presents participation rates by the sex composition of the school, and Figure 5-8 presents participation rates by three categories of location.
Participation rates are lower at co-educational schools compared to single sex schools and at remote schools compared to provincial and metropolitan schools.

5.4.7 State and Sector

Figure 5-9 presents rates of participation by State and Territory, and Figure 5-10 presents rates by sector.
Participation rates were substantially lower in the Northern Territory (NT) compared to the other States and the ACT. A number of factors likely contributed to the lower rates for NT, including a much higher level of remoteness for this jurisdiction, a larger Indigenous population and survey burden for this small jurisdiction as discussed at section 6.7.2. Participation rates at government schools are slightly lower than for non-government schools.
5.5 Summary

The participation rate patterns suggest that there are indeed variations in response to the TIMSS survey across variables related to the outcome measures. Schools with students from lower socio-economic backgrounds have lower response rates, as do schools with lower average performance on the NAPLAN assessment. They suggest that non-response modelling that takes into account socio-economic background and prior performance may be necessary to fully protect survey outcomes from potential biases in survey outcomes.

Larger apparent differences in response patterns were observed with students from the Northern Territory (NT) (Figure 5-9), students from remote locations (Figure 5-8) and students from single sex schools (Figure 5-7). These give some impression of response patterns across Australia as a whole. However, the subpopulation of Australia that will be studied in the following chapters (6 through to 8) is from the Victorian government sector, which has no schools in remote locations, two single sex schools, (and is not the Northern Territory), and so for the purpose of this thesis, these variables were not included in the analysis and simulations.

The following chapter draws upon the investigations of Chapters 4 and 5, about the factors most important in explaining academic performance and response, to examine and compare a range of methods for managing school level non-response under a range of non-response scenarios.
Chapter 6  Weighting adjustments in the presence of school non-response

6.1 Introduction

Reflecting the two stage sample design of most large scale surveys of students – the selection of schools followed by the selection of students within those schools – effective management of unit non-response requires consideration of its effects at each level. As noted in Chapter 3, for surveys conducted under Australia’s National Assessment Plan (NAP), school participation is effectively mandatory, and therefore the major concerns with non-response for these surveys is related to non-response at the student level. However, many student surveys are conducted that do not fall within the NAP, and for these surveys non-response occurs at both student and school levels. Non-response at the school level for these surveys can be quite high. It is likely that the burden of participation in the NAP surveys is leading to higher levels of school non-
response in non-NAP surveys. For example school response rates for a recent report of Child Health and Wellbeing ranged from 33 to 46 percent across the three year levels surveyed, and overall response rates ranged between 9 and 14 percent (Lietz et al., 2015).

In the following chapters, an investigation into the effects of non-response at the school and student levels and an evaluation of a range of non-response adjustment approaches will be conducted. Following from the outcomes of the literature review in Chapter 2, the following non-response adjustment methods will be applied under different non-response scenarios:

- Weighting class adjustments (e.g. non-responding schools within a stratum, non-responding students within a school)
- Post-stratification adjustments (e.g. to match population distribution across categories of performance, socio-economic background and sex)
- Adjustment by estimated response propensity (directly by the inverse of the response propensity, or more broadly by response propensity categories)
- Use of a generalised regression estimator (e.g. making use of correlations between maths performance outcomes and prior performance and socio-economic background)
- Single and multiple imputation using regression models
- Substitution (in the case of school-level non-response)

More detailed information about the adjustments applied will be provided in the relevant chapters.
This chapter evaluates non-response adjustment approaches that can be used to account for school level non-response using simulations based on the response patterns observed in Chapter 5. In Chapter 8, a similar investigation will be conducted with respect to student level non-response. Chapter 7 describes the construction of a simulated student population that will be used for the investigations at the student level.

The ideal dataset to begin with for this study would be data from the full population of students and schools that would give a record for every student for every school in the population. From this database non-response of various kinds could be induced or simulated and a range of non-response management strategies could be applied and compared. Through the annual NAPLAN census tests, a dataset very close to this exists in the Australian educational context. However, this database was not made available for the purposes of this research.

Instead, the following datasets will be used:

- For the investigation of non-response at the school level in this chapter, the ACER School Sampling Frame will be used. This is a database of all of Australia’s schools, updated annually with data supplied to ACER by the Commonwealth and State education systems following the annual census that is conducted by these systems in August each year. This database includes enrolment figures by sex and year level, classifications by location and school type, school contact information and, linked to the postal location of the school, the Index of Education and Occupation (IEO) which is one of the Social Economic Indices for Areas (SEIFA) measures produced the Australian Bureau of Statistics (ABS). Appended to this database for the purposes of this research
is the school mean Year 9 Numeracy score and Numeracy performance bands from the 2010 and 2008 NAPLAN assessments. There are 10 bands which correspond with the first 10 years of formal schooling and are linked to National Minimum Standards, a set of descriptions of the skills and understandings expected at each year level, see ACARA (2017d).

- For the investigation at the student level in chapters 7 and 8, a simulated student level database from a significant subpopulation of Australia’s educational system has been developed from the TIMSS survey data.

6.2 Approach to the investigation

The investigation was limited to the government school sector of Victoria. Victoria is Australia’s second largest State by population, with approximately 5.8 million people. It has a substantial government school sector, educating more than 500,000 students across 13 years of schooling from primary to secondary. At the secondary level, there are 279 schools with Year 9 enrolments.

For each analysis, one thousand samples of 30 schools were selected. The sample size was typical of the school sample sizes drawn from larger jurisdictions (such as the Victorian government sector) for Australian educational surveys.

Schools were sampled with probability proportional to size (PPS), with the assumption that exactly 25 students participated from each sampled school, or the enrolment size if that was less than 25, and that 100% response was obtained from these students. PPS sampling is the most commonly used sampling approach for the selection of schools for Australian student surveys, and the approach used in all NAP surveys.
From each selected sample, a number of schools will be identified as non-respondents under a range of scenarios that will be described in section 6.7, ranging from completely random non-response (MCAR), non-response that is random conditional on survey explanatory variables (MAR), and non-response that is related to outcomes in a manner that cannot be fully explained by the explanatory variables of the survey (NMAR).

A range of methods will then be applied to the responding schools to adjust for the non-responding schools. Population estimates will be produced for each sample and each adjustment method, and outcomes will be compared with estimates that would have been achieved with full response, and with those produced with an adjustment.

In order to isolate the effects to school-level non-response in this chapter the variance associated with the sampling of students within schools will be ignored. The contribution of a school to the population outcomes will be the same as if all students from the school had participated. The effects of student level non-response are examined in Chapter 8.

6.3 School level file preparation

Prior to sampling, the school frame was sorted by variables available on the sampling frame. With systematic sampling of schools using a random start and constant interval and with probability proportional to size selection, the effect of sorting of these variables was to implicitly stratify by these variables. The following is a list of the variables used for stratification:
- School Type. This variable had two categories – Combined (primary and secondary) and Secondary.

- Geographic Location. This is derived from the so-called Geolocation Index, a seven level measure of location first established by MCEECDYA in 2004. Between 2005 and 2016 it was the categorisation of location used in reporting results of Australian national and international educational surveys.

- School-postcode based measure of socio-economic status. This is the Australian Bureau of Statistics SEIFA index of Education and Occupation (IEO) based on the location of the school. For stratification purposes, the within-State deciles generated by ABS were used. The SEIFA score for the school was also available on the frame.

- Year 9 enrolment. Within each substrata formed with the product of school type, location and socio-economic level, schools will be sorted by enrolment size.

This stratification structure is similar to the stratification structure used for the TIMSS survey described in section 4.2.2. With a single State and sector used for this investigation, those were no longer relevant variables. The main difference is the inclusion of the School Type variable, distinguishing between schools with both primary and secondary enrolments, and schools with secondary enrolments only. In addition, a seven level measure of geographic location was used, compared to a three level measure used in TIMSS. Both changes were made because they were considered useful in the selection of potential substitutes for sampled schools, discussed further below (section 6.9.7). The SEIFA decile is used as a school level measure of socio-economic background as this variable is available on the ACER Sampling Frame. The other potential socio-economic background variables discussed in Chapter 4 that might
contribute to explaining achievement such as the Index of Community and Socio-educational Advantage (ICSEA) or one based on data collected of individual students were not available on the sampling frame.

The stratification structure broadly reflects that used in Australian educational surveys, making use of variables available from the ACER Sampling Frame. The structure does not include MySchool variables such as ICSEA or performance means derived from the NAPLAN assessments. Such variables have generally not been made available for sampling for Australian educational surveys, nor for non-response management.

6.4 The outcome measures to be estimated.

The following two outcome measures were examined for the analysis.

1) The mean Year 9 Numeracy score, based on the 2009 NAPLAN measure for the school. The population mean mathematics performance for students from the jurisdiction which is the mean of the school means weighted by their enrolment size, was calculated as 582.5.

2) The proportion of high achieving students defined by those in ‘performance band 10’, which is the top band of NAPLAN scores published by ACARA. The proportion of students in the Victorian government school population in this band is 8.61%.

6.5 A school level measure of prior performance

In Chapter 4 it was observed that, if available, a prior performance measure can contribute substantially to explaining variation in achievement. In explaining the TIMSS Mathematics outcome, the inclusion of both a school level prior measure, the
2010 NAPLAN Year 9 Numeracy mean for the school, and a student level measure, the TIMSS science achievement, added to the explanatory power of the model. The school mean contributed around an additional 5 percentage points to the explanatory power, the student mean of science achievement approximately doubled the explanatory power of the model.

For the analysis in this chapter, the 2010 NAPLAN Numeracy mean for the jurisdiction is the outcome measure being estimated. If a school is identified as a non-respondent under the scenarios to be investigated, the school NAPLAN mean will be missing. The investigation will explore how successfully adjustment methods can account for this missing data in forming an estimate of the mean for the jurisdiction.

The measure of prior performance that will be used in this school level investigation will be the school Year 9 NAPLAN Numeracy mean from two years prior, 2008. This is published on MySchool and is therefore a variable that the statistician responsible for managing non-response could realistically be able to access.

To examine the explanatory power of this variable, it was inserted into regression models explaining the TIMSS Mathematics outcome in line with the Chapter 4 investigation into the variables for explaining achievement. A model using the 2008 NAPLAN mean to explain achievement on the TIMSS measure was compared to Model 2 from that chapter which used the 2010 NAPLAN mean (section 4.4.5, page 145). The comparison is shown in Table 6-1. The two models are very similar, indicating that the 2008 Numeracy mean is almost as effective as the 2010 variable for explaining achievement on the TIMSS outcome.
Table 6-1: Multiple regression on mathematics achievement, TIMSS 2011 Year 8. Comparing performance measures.

<table>
<thead>
<tr>
<th>Parameter Group</th>
<th>Parameter</th>
<th>Source</th>
<th>Models</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>State</td>
<td>NSW (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>VIC</td>
<td>1</td>
<td>x</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>QLD</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>SA</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>WA</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>TAS</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>NT</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>ACT</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Sector</td>
<td>Government (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Catholic</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Independent</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Location</td>
<td>Metro (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Provincial</td>
<td>1</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>School Type</td>
<td>Secondary (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>School Size</td>
<td>Large (reference)</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Composition</td>
<td>Proportion of girls</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>propATSI</td>
<td>1</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>propLBOTE</td>
<td>2</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Socio-Economic</td>
<td>SEIFA IEO</td>
<td>2</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Factors</td>
<td>School ICSEA</td>
<td>2</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Home Educational Resources</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total recurrent funding per student</td>
<td>2</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Sex</td>
<td>Girls</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Boys (reference)</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achievement</td>
<td>NAPLAN Y9 Numeracy school mean (2010)</td>
<td>2</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NAPLAN Y9 Numeracy school mean (2008)</td>
<td>2</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Student achievement in Science (first PV)</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School Climate</td>
<td>Attendance Rate</td>
<td>2</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-sqd</th>
<th>Adj R-sqd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.32</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table models legend

- *x* included in the model but not significant
- *value* standardised regression coefficient (for a significant factor)
- *blank* not included in the model
- *a* a reference category

Given the effectiveness of the 2008 Numeracy mean prior performance measure in explaining variation in outcomes on the TIMSS mathematics outcome (Chapter 4), the
same variable should also perform well as a prior performance measure for predicting the 2010 NAPLAN Numeracy mean for non-responding schools in the investigation in this chapter.

6.6 Explaining achievement using school level data

Chapter 4 involved some regression modelling to investigate which of the variables available from educational survey data best explained achievement in the TIMSS mathematics outcome. The earlier models made use exclusively of school level variables and these combined to explain around 20-30% of variation in achievement. The major factors that contributed to explaining achievement were socio-economic background and prior performance. For example Section 4.4.7 describes ‘Model 2’ where the SEIFA IEO measure and the NAPLAN Numeracy mean along with school location were statistically significant variables in a model that explained 32% of the variation in the TIMSS mathematics outcome. The IEO measures and the NAPLAN mean were the biggest contributors to this model.

Table 6-2 shows a regression using school level variables either available on the ACER Sampling Frame, or readily accessible from MySchool, to explain the 2009 NAPLAN school mean. The models were explored using PROC SURVEYREG and were weighted by the Year 8 enrolment size. Under the Models columns, a blank cell indicates the model did not include the variable; a cell with a period (.) indicates a reference value for a class variable; an ‘x’ indicates a variable included in the model and found not to be significantly different from zero, and numerical values are standardized regression coefficients.
Table 6-2: Multiple regression on mathematics achievement, TIMSS 2011 Year 8. Comparing socio-economic background measures.

<table>
<thead>
<tr>
<th>Parameter Group</th>
<th>Parameter</th>
<th>Models</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td></td>
<td>0.00</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Location</td>
<td>Metro (reference)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provincial</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remote</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School Type</td>
<td>Secondary (reference)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>School Size</td>
<td>Large (reference)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Composition</td>
<td>Proportion of girls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.07</td>
</tr>
<tr>
<td>Socio-Economic Factors</td>
<td>SEIFA IEO</td>
<td></td>
<td>0.68</td>
<td>0.13</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>School ICSEA</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achievement</td>
<td>NAPLAN Y9 Numeracy school mean (2008)</td>
<td></td>
<td>0.93</td>
<td>0.85</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R-sqd</td>
<td></td>
<td>0.46</td>
<td>0.87</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Adj R-sqd</td>
<td></td>
<td>0.46</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The 2008 Numeracy school mean presents as a very strong explanatory variable. On its own, this variable explains 87% of the variation in 2010 Numeracy mean scores. As previously various combinations of the socio-economic factors were explored. The fourth model indicates that, as in the models to explain achievement on the TIMSS survey in Chapter 4, the prior performance measure along with some socio-economic background information were the most important factors. The SEIFA IEO variable appeared to work better in these models compared to the ICSEA measure.

6.7 Non-response scenarios and models

For each sample, a number of schools were identified as non-respondents according to various non-response models, described below. Three scenarios were examined each with increasing rates of non-response. At first, it was assumed that 3 of the 30 sampled
schools were non-respondents, 10% non-response. The second scenario involved 12 non-responding schools (40% non-response) and the third scenario involved 21 non-responding schools (70% non-response).

6.7.1 Missing completely at random non-response

The first model assumed that non-responding schools were missing completely at random (MCAR). For each drawn sample, the sampled schools were allocated a random number. The list of sampled schools was sorted by this random number and the first \( n \) schools \((n = 3, 12, 21)\) were deemed to have not responded. The responding school sizes under each scenario were therefore 27, 18 and 9.

For this model no bias was expected in the unadjusted estimates. It provides a benchmark to examine the effects on standard errors estimated from the different adjustment methods. It can also show if the adjustments lead to unexpected effects on bias.

6.7.2 Missing at random non-response

The second non-response model applied response rates to the sampled schools that were related to the response rates observed (at the student level) in the TIMSS survey across the key measures of performance (as measured by the 2008 NAPLAN Numeracy mean) and socio-economic background (as measured by quintiles of the SEIFA IEO variable) as discussed in section 5.4.3.
The assumption with this approach was that for a survey where school participation was not mandatory, the response rates of schools representing particular populations of students (e.g. from low socio economic backgrounds and with low prior performance) would be proportional to the observed response rates of students from those same backgrounds through the TIMSS survey.

With almost complete school response observed in NAP studies, (where participation is effectively mandatory), there was little available evidence to test this assumption. In practice one could imagine quite unrelated factors leading to school level non-response. The fact for example that the burden of participation falls to larger schools means that especially in smaller jurisdictions, such schools may be particularly less inclined to participate in a voluntary capacity regardless of their socio economic background or prior performance. In their report on improving the response rates of educational surveys in the UK, Sturgis et al. (2006) suggest ‘head teacher’s subjective assessment of pupil competencies’ as a simple example of a variable not otherwise used in addressing school level non response. As discussed in section 2.4.1 Schnepf et al. (2014) found considerable differences in school and pupil patterns of response.

Nevertheless in the absence of available evidence in the Australian context the model of applying response rates by levels of prior performance and socio economic background was considered a reasonable starting point for considering MAR induced non-response at the school level.

To enable comparisons of outcomes across the different non-response models used, three scenarios were tested, with the same fixed number of schools as used in the
MCAR analysis so 3, 12, 21 schools are to be identified as non-respondents. This was achieved through selection of the responding schools by probability proportional to size using the response rate as the measure of size.

**MAR version 1**

In the first MAR model, the TIMSS observed response rates (section 5.4.3) were assigned to each sampled school - according to their prior performance and socio-economic background as measures of size, as shown in Figure 6-1. Thus the measure of size for a school reflects its probability of response. This figure shows the same pattern as the achieved student level response rates shown in section 5.4.3 although the performance categories in Figure 6-1 are based on the 2008 Numeracy mean whereas the performance categories in the earlier figure were based on the 2009 Numeracy mean. Some categories where no data was observed in the TIMSS survey have been added using rates from adjacent categories, for example some of the categories representing low socio-economic quintiles and high performance at the top right. This was necessary in case data from those categories occur in the population of schools being used for sampling in this chapter.
Prior to sampling the sampled schools were randomly sorted and the desired number of schools were sampled systematically with probability proportional to size based on the assigned response rate shown in Figure 6-1. In the first scenario for example, 27 schools were sampled leaving 3 non-responding schools. The random sorting of schools prior to sampling was so that stratification effects did not interfere with the intended response distribution. Alternatively, schools could have been sampled with probability proportional to size but without a systematic sampling approach.

With schools subject to non-response, the achieved response rates could vary considerably from sample to sample under this approach. With six categories of prior performance – defined against the 10th, 25th, 50th, 75th and 90th percentiles – and quintiles of the IEO measure used to classify against socio-economic background, the response rates were distributed across 30 separate categories in the cross classification. Some subpopulations defined by prior performance and socio economic background
have very small numbers of schools, and with 30 schools sampled, it was common to
have no schools sampled in some subpopulations.

All sampled schools are representing the same number students in the population – the
sampling interval, the total Year 9 enrolment across the 279 schools in the population
divided by the number of schools sampled. So when a school does not respond, that
number of students within the subpopulation within that sampling interval are not being
represented, and the intention of the non-response management methods is to address
this. The expected value of the response rates, weighted to reflect the number of
students being represented by the students in the participating schools, would be
expected to be in line with the response rates used as measures of size. The average
response rate weighted to the number of students in the population for this MAR
version was around 88%, the same response rate observed with the TIMSS data.
However, for an individual sample, response rates for a school within a particular
subpopulations might easily be 0%, 100% or something in between.

**MAR version 2**

It was desired to also test the non-response management approaches in an environment
with lower response rates but still under a situation where the non-response could be
explained by prior performance and socio economic background, that is, with non-
response that was MAR. The second MAR scenario began with the same response rates
that were used in MAR1, but added a further drop in response rates for successively
lower categories of prior performance. Response rates for schools in the highest decile
on the 2008 Numeracy mean were 95% of the rates used in MAR1, and this dropped a
further 5% for each category, down to a factor of 65% for schools in the lowest decile.
For example, schools in the highest decile based on the 2008 Numeracy mean and in the
fifth SEIFA quintile had a response rate in MAR1 of 93% (based on TIMSS outcomes). In MAR2, this dropped to $0.93 \times 0.95 = 0.88$. Rates used in MAR1 across all IEO quintiles in this highest Numeracy decile dropped by this factor of 0.95. At the other end of the scale, all schools in the lowest Numeracy 2008 decile had response rates drop by a further 35%. Schools in the first quintile of the lowest Numeracy category had a response rate of 78% under MAR1, but this dropped to $0.78 \times 0.65 = 0.51$ under MAR2.

Figure 6-2 charts the assumed measures of size used in this second model investigating missing at random responses at the school level. It shows a complex interaction between response rates, performance and socio-economic background which was an observation from response rates from the TIMSS analysis.

![Figure 6-2: Measures of size (response probability) by 2008 Numeracy mean x IEO quintile: School non-response model 'MAR2'](image)
6.7.3 Non-response related to outcomes

The third part of the investigation will be to examine non-response management approaches when some of the missing data is related to survey outcomes but which cannot be explained by other survey variables, so the data are NMAR, not missing at random.

Conditional on the response rates by categories defining levels of the 2008 NAPLAN school Numeracy mean and quintiles of IEO that were used in the first missing at random (MAR1) algorithm described above, a further factor relating to levels of the outcome variable, the 2009 Numeracy mean was added. Response rates for schools in the top quartile on the 2009 Numeracy mean were dropped by a further 10% compared to the response rates assumed under MAR. The response rate dropped a further 20% for each successive quartile.

For example, the response rates for students from schools between the 50th and 75th percentiles based on the 2008 Numeracy mean, and in the fourth IEO quintile had a response rate under MAR2 of 86%. Those schools in the top quartile based on the 2009 Numeracy mean had a revised response rate of 0.86 * 0.9 = 78%. Those in the next quartile had a revised response rate of 0.86 * 0.7 = 61%. Students from schools in the lowest category based on the 2008 Mean and the lowest IEO quintile, and in the lowest quartile based on the 2009 mean had a response rate of 23%. Figure 6-3 displays the assumed response rates, with each line of the chart representing quartiles of the 2009 Numeracy mean.
As with the MAR models, the assumed response rates were assigned to each school as measures of size, and fixed numbers of schools – 27, 18 and 9 - were sampled from this list with probability proportional to size.

### 6.8 Expected values of response rates across prior performance and IEO

Figure 6-4 to Figure 6-7 compare, for each non-response model and for each level of school response (n = 27, 18, 9) the response rates estimated from the average response for the 1000 samples with the measure of size that was used in the algorithm to induce non-response. We would expect the response rate to mirror the measure of size, although the level will depend on the amount of non-response that was set.
The main observation from the charts is that the pattern of the expected value of the response rates generally compares well with the measure of size rate, although there may be some distance between the two lines. The measure of size induced response rates for each category of prior performance x IEO relative to the other prior performance x IEO categories, but the size of the non-response was a function of the number of schools set as non-responding in each scenario.

The charts provide confirmation that the method for inducing non-response for this component of the research functioned as intended.

While the average rate over 1000 samples fits with the relative measures of size, the rates achieved from sample to sample may vary very widely from that average. With fewer than 30 schools sampled across the 30 categories of performance by socio-economic background, this is not surprising. The count of responding schools compared to the number sampled within a category could vary widely for any particular sample, from zero schools in the category through to all schools, and anything in between.
In the MCAR scenarios (Figure 6-4), the response rate patterns are essentially straight lines, corresponding to the fact that the measure of size (MOS) used in the algorithm for generating non-response was constant for each category, set (arbitrarily) at 0.5. The average response rates over the 1000 samples drawn show approximately equal probability as expected. Occasional small variations occur in categories with very small numbers of schools and students. The lines are located at response rate levels of 0.9, 0.6 and 0.3, reflecting the proportion of responding schools: 27, 18 and 9 schools from 30.
In the first MAR scenario, (MAR1 - Figure 6-5), the pattern of the expected value of the response rates across the categories is consistent with the MOS used in the non-response algorithm for this model, which corresponds with the pattern of response observed in the TIMSS survey as shown in Figure 5-4.

Figure 6-6: Expected value of the response rate used in the MAR2 non-response model across performance x socio-economic background categories
In Figure 6-6, with a lower overall response rate, the average rates per category across the 1000 samples are lower than the measures of size used in the model – however the pattern is consistent with that model.

![Figure 6-7: Expected value of the response rate used in the NMAR non-response model across performance x socio-economic background categories](image)

In Figure 6-7 the average MOS across the four Numeracy quartiles used in the NMAR non-response algorithm (Figure 6-3) is displayed, and compared with the average response rates for each scenario.

### 6.9 Non-response management approaches

For each model and scenario described above, the following strategies for managing school level non-response will be investigated.

#### 6.9.1 No school adjustment

This approach made the assumption that the missing schools are missing completely at random. This is also equivalent to a weighting adjustment applied to all schools equal to
the ratio of the number sampled and the number participating, as altering the design weight by a constant factor does not affect the sample estimate.

6.9.2 Adjustment by weighting classes within the stratum
Weighting class adjustments were limited because of the small sample sizes, as well as in some cases the small responding school size (as few as 9 schools). It was decided to limit the number of weighting classes to two so that a reasonable number of responding schools would be expected within each weighting class, avoiding an excessive degree of collapsing of strata across the 1000 samples. This ensured better ‘like for like’ comparisons of non-response adjustments.

Three different weighting class adjustments were investigated:

- Adjustments by high and low IEO, defined against the median IEO score;
- Adjustments by high and low prior performance, defined against the median 2008 NAPLAN Numeracy mean;
- Adjustments by location, metropolitan versus non-metropolitan.

An overall non-response adjustment weight equal to the ratio of the number of sampled schools and the number that participated will be made within each group separately. This is the same type of school level weighting adjustment used in TIMSS, as described in section 4.3.1.

6.9.3 Response propensity estimation weight
In this approach the propensity of school response was estimated using a logistic regression model, with the probability of response predicted by the NAPLAN 2008 Numeracy mean and the IEO score for the school. The logistic regression was weighted
by the school design weight. The inverse of the estimated response probability was calculated and this weight was applied in the estimation of outcomes.

6.9.4 Adjustment by weighting classes defined by response propensity

Schools were classified into high and low response propensity classes defined against the median propensity, and weighting class adjustments as described in section 6.9.2 were applied.

6.9.5 A regression estimator using the 2008 Numeracy mean and the IEO score

Auxiliary data known to be correlated with achievement, the 2008 Numeracy mean and the IEO measure, and available for all sampled cases, respondents and non-respondents, were used to adjust the design based estimate of the 2009 Numeracy mean in the manner described in section 2.7.4. A Horvitz-Thompson estimator of the mean mathematics score was generated using the response data, and an adjustment to this mean was applied equal to the product of the estimated vector of regression coefficients based on the response data multiplied by the difference between sample based estimate of auxiliary totals and the response based estimate of those totals.

6.9.6 Regression based imputation

Using the participating sample, a regression of the outcome measure against the stratification variables used in sampling as described in section 6.3 (School Type, Location, and the IEO score), as well as the 2008 Numeracy mean were used to predict the 2009 Numeracy school mean score that would have been obtained from the non-participating schools. The population mean was then estimated as the weighted mean of
the school means of the participating schools, and the imputed school means of the non-participating schools.

6.9.7 School substitution

Schools adjacent to the sampled school on the frame were allocated as possible substitutes. As noted above prior to sampling the school sampling frame was organised (sorted) by the stratification variables School Type, Location, Socio-economic level, as well as school size and so adjacent schools were generally similar to the sampled school for these variables. Normally the school on the frame immediately below the sampled school was assigned as the substitute. In cases where the defined substitute crossed implicit stratum boundaries, the assignment may be altered. For example if the last school in the stratum defined by secondary metropolitan schools from the highest socio-economic area is sampled (a relatively small school), then the school immediately below may be a larger school from a lower socio-economic area, but the school above may be similar in size and from the same socio-economic level. In this case it would be assigned as the substitute. This approach is a method applied in all NAP surveys and might be described as a version of ‘near neighbour’ imputation.

6.10 Presentation of outcomes

For each analysis, the estimates of the outcomes from the 1000 selected samples were averaged to generate an estimate of the expected value of the statistic under the assumed non-response model. Estimates of the corresponding standard error were also obtained from the 1000 samples. These were tabulated for each model and each scenario and are presented in detail in section 6.12. In addition, boxplots were prepared to show the
distribution of outcomes across the samples for each model, and these accompany the summary tables of expected values and standard errors in section 6.13.

The default boxplot presentation from SAS PROC SGPLOT was used to chart the boxplots. The components of each boxplot are summarised in the image below and accompanying text, extracted from the SAS documentation.

![Parts of a box plot (SAS output)](image)

The bottom and top edges of the box indicate the intra-quartile range (IQR). That is, the range of values between the first and third quartiles (the 25th and 75th percentiles). The marker inside the box indicates the mean value. The line inside the box indicates the median value.

By default, the whiskers that extend from each box indicate the range of values that are outside of the intra-quartile range. However, the values are close enough not to be considered outliers (a distance less than or equal to 1.5*IQR).

Any points that are a distance of more than 1.5*IQR from the box are considered to be outliers. By default, these points are indicated by markers.

Figure 6-8: Box plot description from (SAS/GRAPH(R) 9.2: Statistical Graphics Procedures Guide).
With two statistics estimated over four non-response models, and three levels of response, a total of 24 outcome summaries were prepared. A selection of those outcomes that illustrate the main observations of the effects of different patterns of non-response and the adjustment approaches used to manage these are presented in more detail. Following that summaries of all investigations are presented in graphical and tabular form in section 6.13. The tables presented include statistics reflecting the bias and precision of each adjustment method under each non-response model as discussed further below.

6.11 Bias and precision statistics

Following the presentation of selected outcomes in 6.12, the summary of all outcomes (section 6.13) includes additional summary statistics on the precision and bias of estimates from the respective models and methods. The precision and bias summaries (Table 6-9 and Table 6-10) include the following statistics:

- The estimate of the expected value of the population mean and standard error for each management method;
- The bias, the difference between the expected value of the estimated mean following each investigation and the population mean:

\[
Bias = E(\bar{y}) - \bar{Y}
\]
• The mean square error (MSE), the expected value of the squared difference between the estimated mean for a sample and the population mean, equal to the sum of the variance and the squared bias:

\[ MSE = E[\bar{y} - \bar{Y}]^2 = Bias^2 + Variance \]  

\[ Variance = E[\bar{y} - E(\bar{y})]^2 \]  

The mean square error (MSE) gives the average squared difference of each estimate to the population parameter.

• The root mean square error (RMSE), the square root of MSE:

\[ RMSE = \sqrt{MSE} \]  

Taking the square root to form the root mean square error (RMSE) provides an indication of this distance between the estimated value and the population value in the units of the outcome variable (in this case the scores on the Numeracy assessment).

6.12 Selected results

6.12.1 Introduction

The following is an extract of six of the 24 analyses comparing school non-response adjustment approaches across the different non-response mechanisms and response rate scenarios. These are intended to highlight key observations across the full set of analyses. Considerations when evaluating the estimates for each of the adjustment
methods included how closely the distribution of estimates was centred on the population parameter being estimated, the shape of the distribution of estimates – whether symmetric or skewed – and the number and degree of outlier estimates the adjustment method generated. The variation in estimated outcomes are summarised in the standard error statistics in the tables, as well as in the shapes of the boxplots, including the interquartile range as displayed in the width of the boxplots and the overall range including length of the tails and outlying values.

6.12.2 MCAR, low response, estimating the mathematics mean

Figure 6-9 shows the distribution of outcomes of the various adjustment methods with low response levels (18 responding schools out of 30) in estimating the mathematics mean when missing data was MCAR. Table 6-3 below the box plots summarises each outcome with the expected value and standard error (the mean and standard deviation of the 1000 estimates) for each management approach. The population estimate was 582.5. The box plots have labels a – j representing the non-response management methods applied. These labels are described in the accompanying Table 6-3, which gives the means and standard errors.
Figure 6-9: Distribution of estimated mathematics mean under MCAR school non-response with 18 respondents

Table 6-3: Expected value and standard error of Numeracy mean under MCAR school non-response with 18 respondents

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full response</td>
<td>582.59</td>
<td>4.76</td>
</tr>
<tr>
<td>b</td>
<td>Respondents - no adjustments</td>
<td>582.63</td>
<td>6.24</td>
</tr>
<tr>
<td>c</td>
<td>High / Low IEO adjustment</td>
<td>582.65</td>
<td>6.19</td>
</tr>
<tr>
<td>d</td>
<td>High / Low performance adjustment</td>
<td>582.60</td>
<td>6.14</td>
</tr>
<tr>
<td>e</td>
<td>Metro / non metro adjustment</td>
<td>582.69</td>
<td>6.37</td>
</tr>
<tr>
<td>f</td>
<td>Response propensity</td>
<td>581.38</td>
<td>4.76</td>
</tr>
<tr>
<td>g</td>
<td>High / Low propensity adjustment</td>
<td>582.19</td>
<td>5.77</td>
</tr>
<tr>
<td>h</td>
<td>Regression estimator (auxiliaries IEO and 2008_Nmean)</td>
<td>582.41</td>
<td>5.10</td>
</tr>
<tr>
<td>i</td>
<td>Regression imputation: 2009_Nmean = IEO, 2008_Nmean, school type, location</td>
<td>582.41</td>
<td>5.08</td>
</tr>
<tr>
<td>j</td>
<td>School substitution</td>
<td>582.64</td>
<td>4.85</td>
</tr>
</tbody>
</table>
Most, but not all approaches to managing the non-response of 12 of the 30 schools produce outcomes that are centred on the population mean. This was similar across the scenarios investigated with varying levels of non-response. The distribution of the adjustment on the basis of classes of response propensity is centred slightly below the population mean, as is the method involving adjustment on the basis of propensity weighting classes. Those methods did not perform well with low (12/30) or very low (21/30) numbers of non-responding schools, but performed reasonably well when the non-response was less (3/30). Of the weighting class adjustments, the adjustment based on location is the least precise, as was expected given that location was not a factor in the non-response mechanism. The standard errors are around 20% higher for the weighting class adjustments compared to the estimate when there is full response. The interquartile ranges are wider for these methods, and there are more examples of outlying estimates. In general, the weighting class adjustments had more variable outcomes, (larger standard errors). The regression based approaches, the regression estimator and the regression imputation both performed well with respect to bias and precision. School substitution performed even better with a distribution centred on the mean and no loss of precision.

6.12.3 MCAR, very low response, estimating the proportion of top students

Figure 6-10 shows the distribution of outcomes of the various adjustment methods with very low response levels (9 responding schools out of 30) in estimating the proportion of students in the top band for numeracy. Table 6-4 below the box plots summarises each outcome with the expected value and standard error.
While the distributions of most of the adjustment methods were centred on the mean, most produced very skewed distributions of outcomes with many individual estimates a
considerable degree higher than the population estimate of 8.61% - sometimes more than double or even triple the population figure - pushing the mean estimate in some cases a long way from the median. Particularly with respect to the weighting class adjustment methods, the median estimate for these distributions, probably a more relevant descriptive estimate of the distribution centre in these skewed cases, was 1-2 percentage points below the population value. Once again the response propensity adjustment underestimated the population mean by quite a long distance. The regression estimator was slightly biased downwards, as was the method involving regression imputation. As with the approaches investigating the school mean, school substitution approach stands out as the best of the adjustment methods, with precise estimates and minimal bias.

6.12.4 MAR1, very low response estimating the mathematics mean

Figure 6-11 compares distributions under the different adjustment methods under very low response and data that was MAR. Table 6-5 shows estimated means and standard errors under each adjustment method.
Figure 6-11: Distribution of estimated mathematics mean under MAR (v1) school non-response with 9 respondents.

Table 6-5: Expected value and standard error of Numeracy mean under MAR (v1) school non-response with 9 respondents

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full response</td>
<td>582.59</td>
<td>4.76</td>
</tr>
<tr>
<td>b</td>
<td>Respondents - no adjustments</td>
<td>583.33</td>
<td>7.95</td>
</tr>
<tr>
<td>c</td>
<td>High / Low IEO adjustment</td>
<td>582.91</td>
<td>7.84</td>
</tr>
<tr>
<td>d</td>
<td>High / Low performance adjustment</td>
<td>582.57</td>
<td>7.54</td>
</tr>
<tr>
<td>e</td>
<td>Metro / non metro adjustment</td>
<td>583.65</td>
<td>8.48</td>
</tr>
<tr>
<td>f</td>
<td>Response propensity</td>
<td>579.78</td>
<td>4.99</td>
</tr>
<tr>
<td>g</td>
<td>High / Low propensity adjustment</td>
<td>581.51</td>
<td>6.75</td>
</tr>
<tr>
<td>h</td>
<td>Regression estimator (auxiliaries IEO and 2008_Nmean)</td>
<td>581.21</td>
<td>5.74</td>
</tr>
<tr>
<td>i</td>
<td>Regression imputation: 2009_Nmean = IEO,2008_Nmean, school type, location</td>
<td>581.97</td>
<td>5.43</td>
</tr>
<tr>
<td>j</td>
<td>School substitution</td>
<td>582.90</td>
<td>4.61</td>
</tr>
</tbody>
</table>

Once again, some quite skewed distributions of estimates in this scenario with low response rates and missing at random data. Most approaches have distributions centred on, or at least close to the population mean. Both the response propensity approaches
again underestimate the mean. Larger interquartile ranges are again apparent for the weight adjustment approaches compared to the regression and imputation approaches. Weighting class adjustments by IEO and by location performed little better – and through producing estimates even further away than the population mean - arguably worse than the approach with no adjustments. The weighting class adjustment by performance level recaptures the population mean more successfully than the other weight adjustment approaches, which is to be expected as it incorporates a major factor relating to the generation of missing data in this model. The substitution stands out again as the most successful approach both with respect to minimal bias and good precision. The regression estimator and the regression imputation also have estimates close to the population mean, with minimal loss of precision.
6.12.5 MAR 2, low response estimating the proportion in the top Numeracy band

Figure 6-12 displays distributions of results when estimating the proportion in the top numeracy band under the stronger missing at random scenario (MAR2) and with low response.

Figure 6-12: Distribution of estimated proportion in top Numeracy band under MAR (v2) school non-response with 18 respondents
Table 6-6: Expected value and standard error of proportion in top Numeracy band under MAR (v2) school non-response with 18 respondents

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full response</td>
<td>8.64</td>
<td>1.96</td>
</tr>
<tr>
<td>b</td>
<td>Respondents - no adjustments</td>
<td>10.23</td>
<td>2.62</td>
</tr>
<tr>
<td>c</td>
<td>High / Low IEO adjustment</td>
<td>9.55</td>
<td>2.32</td>
</tr>
<tr>
<td>d</td>
<td>High / Low performance adjustment</td>
<td>9.12</td>
<td>2.41</td>
</tr>
<tr>
<td>e</td>
<td>Metro / non metro adjustment</td>
<td>10.09</td>
<td>2.63</td>
</tr>
<tr>
<td>f</td>
<td>Response propensity</td>
<td>8.34</td>
<td>1.81</td>
</tr>
<tr>
<td>g</td>
<td>High / Low propensity adjustment</td>
<td>9.05</td>
<td>2.15</td>
</tr>
<tr>
<td>h</td>
<td>Regression estimator (auxiliaries IEO and 2008_Nmean)</td>
<td>8.03</td>
<td>1.79</td>
</tr>
<tr>
<td>i</td>
<td>Regression imputation: 2009_Nmean = IEO , 2008_Nmean, school type, location</td>
<td>8.64</td>
<td>1.94</td>
</tr>
<tr>
<td>j</td>
<td>School substitution</td>
<td>8.75</td>
<td>1.82</td>
</tr>
</tbody>
</table>

Adjustments by weighting classes defined by IEO and location remain biased with distributions centred above the population proportion. Once again the weighting class adjustment by classes of performance is less biased, although is less precise than the propensity score and regression based adjustments, all of which do quite well in addressing the MAR non-response without substantially inflating the standard errors.

6.12.6 NMAR, moderate response estimating the Numeracy mean

Figure 6-13 compares distributions of results of the adjustment methods when data was NMAR, with moderate levels of response. Table 6-7 compares estimated means and standard errors for each approach.
Figure 6-13: Distribution of estimated mathematics mean under NMAR school non-response with 27 respondents

Table 6-7: Expected value and standard error of Numeracy mean under NMAR school non-response with 27 respondents

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full response</td>
<td>582.59</td>
<td>4.76</td>
</tr>
<tr>
<td>b</td>
<td>Respondents - no adjustments</td>
<td>587.03</td>
<td>5.50</td>
</tr>
<tr>
<td>c</td>
<td>High / Low IEO adjustment</td>
<td>582.66</td>
<td>7.99</td>
</tr>
<tr>
<td>d</td>
<td>High / Low performance adjustment</td>
<td>574.65</td>
<td>10.32</td>
</tr>
<tr>
<td>e</td>
<td>Metro / non metro adjustment</td>
<td>587.43</td>
<td>6.50</td>
</tr>
<tr>
<td>f</td>
<td>Response propensity</td>
<td>583.12</td>
<td>4.72</td>
</tr>
<tr>
<td>g</td>
<td>High / Low propensity adjustment</td>
<td>570.64</td>
<td>8.72</td>
</tr>
<tr>
<td>h</td>
<td>Regression estimator (auxiliaries IEO and 2008_Nmean)</td>
<td>570.53</td>
<td>9.88</td>
</tr>
<tr>
<td>i</td>
<td>Regression imputation: 2009_Nmean = IEO, 2008_Nmean, school type, location</td>
<td>582.91</td>
<td>4.79</td>
</tr>
<tr>
<td>j</td>
<td>School substitution</td>
<td>583.73</td>
<td>4.66</td>
</tr>
</tbody>
</table>
This scenario involving missing data relating to survey variables as well as to the Numeracy outcome shows that most methods have been unable to completely remove the bias arising from non-response, even with reasonably good response rates (27 out of 30 schools). This is to be expected as all methods assume some form of MAR. The response propensity adjusted estimate in this case is centred closest to the population mean, although the other regression based approaches are also close. Once again the weighting class adjustment methods do least well in reducing bias, and have the largest standard errors.

6.12.7 NMAR, very low response estimating the proportion in the top Numeracy band

Figure 6-14 compares distributions of estimates of the proportion of students in the top numeracy band, across the different adjustment methods under NMAR with very low school response.
Figure 6-14: Distribution of estimated proportion in top Numeracy band under NMAR school non-response with 9 respondents

Table 6-8: Expected value and standard error of proportion in top Numeracy band under NMAR school non-response with 9 respondents

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full response</td>
<td>8.64</td>
<td>1.96</td>
</tr>
<tr>
<td>b</td>
<td>Respondents - no adjustments</td>
<td>10.19</td>
<td>3.71</td>
</tr>
<tr>
<td>c</td>
<td>High / Low IEO adjustment</td>
<td>9.45</td>
<td>3.30</td>
</tr>
<tr>
<td>d</td>
<td>High / Low performance adjustment</td>
<td>9.13</td>
<td>3.38</td>
</tr>
<tr>
<td>e</td>
<td>Metro / non metro adjustment</td>
<td>10.16</td>
<td>3.89</td>
</tr>
<tr>
<td>f</td>
<td>Response propensity</td>
<td>7.75</td>
<td>1.79</td>
</tr>
<tr>
<td>g</td>
<td>High / Low propensity adjustment</td>
<td>8.65</td>
<td>2.97</td>
</tr>
<tr>
<td>h</td>
<td>Regression estimator (auxiliaries IEO and 2008_Nmean)</td>
<td>6.88</td>
<td>2.09</td>
</tr>
<tr>
<td>i</td>
<td>Regression imputation: 2009_Nmean = IEO,2008_Nmean, school type, location</td>
<td>8.53</td>
<td>2.22</td>
</tr>
<tr>
<td>j</td>
<td>School substitution</td>
<td>8.77</td>
<td>1.94</td>
</tr>
</tbody>
</table>

The weighting class adjustments have performed poorly under these conditions with non-response associated with the numeracy measure and very low response rates. The expected value of the proportion is generally well above the population value, and the
distributions are very skewed with estimates for some samples a long way from the population proportion. Regression imputation and school substitution have both performed well under these conditions with minimal bias and minimal inflation of standard errors.

6.13 Outcomes summaries

6.13.1 Distributions of means, precision and bias

Figure 6-15 and Table 6-9 on the following pages summarise the results of all the school non-response investigations undertaken with charts showing the distributions under each model and non-response level and tables with statistics summarising the effects of bias and precision for each scenario. Looking across the rows of Figure 6-15, the effects of lower response rates on width of boxplots representing the interquartile range is clear. Also clear is the increase in the number of outlying estimates, both as response declines and the drivers of non-response become more complex. Looking down the column, it is clear that the more complex the factors underlying non-response, the bigger the variation in performance across the different non-response management approaches, even with relatively high rates of school response. Another clear feature is the instability of estimates under the more complex scenarios, with very skewed distributions with the mean being a less effective measure of the centre of distributions. As discussed in the more detailed individual summaries, the regression based methods and school substitution stand out as better performing approaches across the scenarios.
Figure 6-15: Summary distributions – estimating the population mean – school level non-response
Table 6-9: Precision and bias. Estimates of population mean. All non-response models

226


Table 6-9 quantifies the combined impact of precision and bias across the weighting management approaches. In the MCAR scenarios there is minimal bias. The larger standard errors and relative standard errors when non-response is low clearly contrast with the standard errors when response is higher. Even under the MAR scenarios, the bias remains low and the effects on standard errors due to lower rates of response overall are clear. Variations in the effectiveness of the management methods employed are also evident. A good summary indicator of relative effectiveness under the different scenarios is the size of RMSE. As the non-response rates drop and as the complexity of non-response increases, the RMSEs for those methods that do incorporate most or all of the factors underlying that non-response remain fairly stable. In contrast, the factors that do not take full account of factors such as prior performance and socio-economic background exhibit much larger RMSEs, sometimes three times the size compared to the full-response scenario.

With low response rates and data missing at random, the relative effectiveness of some management methods – particularly substitution and the regression based approaches - over others becomes clearer. The introduction of a component of missingness not explained by other survey variables in the NMAR models leads to biased outcomes across all methods, but with the best performing methods in the MAR models still clearly performing relatively better in this situation.

6.13.3 Distributions of proportions, precision and bias
Figure 6-16 and Table 6-10 show the same summary charts and tables with respect to the other outcome variable explored, the proportion of students in the top numeracy band. More than was observed for the distribution of mean estimates, the effect on estimates of the proportion as non-response drops tends to be the incidence of a higher
number of more extreme observations, leading to more skewed distributions. When response was very low, the expected values of adjustment approaches were close to the population value, but well over half of the estimates (as represented by the median) were below that value. This occurred even under the MCAR scenario. School substitution and regression based approaches were clearly better overall.
Figure 6-16: Summary distributions – estimating the population proportion – school level non-response
Table 6-10: Precision and bias. Estimates of proportion. All non-response models

<table>
<thead>
<tr>
<th></th>
<th>MCAR 27</th>
<th></th>
<th>MCAR 18</th>
<th></th>
<th>MCAR 9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EV</td>
<td>SE</td>
<td>bias</td>
<td>MSE</td>
<td>RMSE</td>
<td>EV</td>
</tr>
<tr>
<td>a</td>
<td>8.64</td>
<td>1.96</td>
<td>0.03</td>
<td>3.84</td>
<td>1.96</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>8.64</td>
<td>2.09</td>
<td>0.03</td>
<td>4.36</td>
<td>2.09</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>8.65</td>
<td>2.09</td>
<td>0.04</td>
<td>4.36</td>
<td>2.09</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>8.63</td>
<td>2.08</td>
<td>0.02</td>
<td>4.32</td>
<td>2.08</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>8.65</td>
<td>2.09</td>
<td>0.04</td>
<td>4.36</td>
<td>2.09</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>8.51</td>
<td>1.96</td>
<td>-0.10</td>
<td>3.83</td>
<td>1.96</td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>8.52</td>
<td>2.01</td>
<td>-0.09</td>
<td>4.04</td>
<td>2.01</td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>8.60</td>
<td>1.99</td>
<td>-0.01</td>
<td>3.94</td>
<td>1.99</td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>8.72</td>
<td>2.02</td>
<td>0.11</td>
<td>4.10</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>j</td>
<td>8.64</td>
<td>1.94</td>
<td>0.03</td>
<td>3.76</td>
<td>1.94</td>
<td></td>
</tr>
</tbody>
</table>

MAR1 27

<table>
<thead>
<tr>
<th></th>
<th>MAR1 18</th>
<th></th>
<th>MAR1 9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EV</td>
<td>SE</td>
<td>bias</td>
<td>MSE</td>
</tr>
<tr>
<td>a</td>
<td>8.64</td>
<td>1.96</td>
<td>0.03</td>
<td>3.84</td>
</tr>
<tr>
<td>b</td>
<td>8.87</td>
<td>2.09</td>
<td>0.25</td>
<td>4.42</td>
</tr>
<tr>
<td>c</td>
<td>8.76</td>
<td>2.01</td>
<td>0.15</td>
<td>4.05</td>
</tr>
<tr>
<td>d</td>
<td>8.81</td>
<td>2.08</td>
<td>0.02</td>
<td>4.32</td>
</tr>
<tr>
<td>e</td>
<td>8.66</td>
<td>2.09</td>
<td>0.03</td>
<td>4.36</td>
</tr>
<tr>
<td>f</td>
<td>8.50</td>
<td>1.95</td>
<td>-0.02</td>
<td>3.80</td>
</tr>
<tr>
<td>g</td>
<td>8.54</td>
<td>1.94</td>
<td>0.07</td>
<td>3.75</td>
</tr>
<tr>
<td>i</td>
<td>8.73</td>
<td>1.98</td>
<td>0.12</td>
<td>3.93</td>
</tr>
<tr>
<td>j</td>
<td>8.60</td>
<td>1.94</td>
<td>-0.01</td>
<td>3.76</td>
</tr>
</tbody>
</table>

MAR2 27

<table>
<thead>
<tr>
<th></th>
<th>MAR2 18</th>
<th></th>
<th>MAR2 9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EV</td>
<td>SE</td>
<td>bias</td>
<td>MSE</td>
</tr>
<tr>
<td>a</td>
<td>8.64</td>
<td>1.96</td>
<td>0.03</td>
<td>3.84</td>
</tr>
<tr>
<td>b</td>
<td>9.42</td>
<td>2.10</td>
<td>0.81</td>
<td>5.08</td>
</tr>
<tr>
<td>c</td>
<td>9.05</td>
<td>1.93</td>
<td>0.43</td>
<td>3.92</td>
</tr>
<tr>
<td>d</td>
<td>8.72</td>
<td>2.00</td>
<td>0.11</td>
<td>4.02</td>
</tr>
<tr>
<td>e</td>
<td>9.39</td>
<td>2.16</td>
<td>0.78</td>
<td>5.28</td>
</tr>
<tr>
<td>f</td>
<td>8.56</td>
<td>1.84</td>
<td>-0.05</td>
<td>3.39</td>
</tr>
<tr>
<td>g</td>
<td>8.71</td>
<td>1.90</td>
<td>0.10</td>
<td>3.63</td>
</tr>
<tr>
<td>h</td>
<td>8.21</td>
<td>1.79</td>
<td>-0.40</td>
<td>3.36</td>
</tr>
<tr>
<td>i</td>
<td>8.67</td>
<td>1.91</td>
<td>0.06</td>
<td>3.67</td>
</tr>
<tr>
<td>j</td>
<td>8.64</td>
<td>1.84</td>
<td>0.03</td>
<td>3.38</td>
</tr>
</tbody>
</table>

NMAR 27

<table>
<thead>
<tr>
<th></th>
<th>NMAR 18</th>
<th></th>
<th>NMAR 9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EV</td>
<td>SE</td>
<td>bias</td>
<td>MSE</td>
</tr>
<tr>
<td>a</td>
<td>8.64</td>
<td>1.96</td>
<td>0.03</td>
<td>3.84</td>
</tr>
<tr>
<td>b</td>
<td>10.17</td>
<td>2.38</td>
<td>1.56</td>
<td>8.06</td>
</tr>
<tr>
<td>c</td>
<td>9.46</td>
<td>2.06</td>
<td>0.85</td>
<td>4.95</td>
</tr>
<tr>
<td>d</td>
<td>9.50</td>
<td>2.39</td>
<td>0.89</td>
<td>6.48</td>
</tr>
<tr>
<td>e</td>
<td>10.07</td>
<td>2.40</td>
<td>1.45</td>
<td>7.85</td>
</tr>
<tr>
<td>f</td>
<td>8.93</td>
<td>1.98</td>
<td>0.32</td>
<td>4.01</td>
</tr>
<tr>
<td>g</td>
<td>9.12</td>
<td>2.05</td>
<td>0.51</td>
<td>4.45</td>
</tr>
<tr>
<td>h</td>
<td>8.44</td>
<td>1.85</td>
<td>-0.17</td>
<td>3.49</td>
</tr>
<tr>
<td>i</td>
<td>8.94</td>
<td>1.94</td>
<td>0.33</td>
<td>3.88</td>
</tr>
<tr>
<td>j</td>
<td>8.93</td>
<td>1.87</td>
<td>0.32</td>
<td>3.61</td>
</tr>
</tbody>
</table>
6.14 Summary of outcomes to approaches for managing non-response

The results from the investigations into non-response management approaches under various non-response scenarios demonstrated some clear overall observations:

1) Weighting class adjustments that did not take into account the drivers of non-response were less successful in managing non-response than those that partially or fully addressed those factors. They were generally more biased and less precise. For example adjustments by weighting classes defined by location consistently did less well than adjustments within weighting classes defined by performance measures. In turn, those were less successful than the methods that took account of both socio-economic background and performance.

2) Weighting class adjustments, which by necessity were quite crude, with just two weighting classes defined per variable, were consistently not as successful as regression based approaches, where adjustments were more fine-grained.

3) While most adjustments were able to produce expected values close to the population values under MCAR and MAR conditions - the distributions of some adjustments, particularly the simple weighting class adjustments – were very skewed with many individual samples producing estimates a long way from the population mean. In some cases the adjustments appeared to perform no better, and arguably worse than when no adjustment was made.

4) As expected, none of the methods were completely able to remove the bias that was directly related to the outcome measure.

5) The response propensity approach tended to have good precision, but outcomes were sometimes quite biased. With a relatively small number of data points in
order to model response propensity – sometimes as few as 9 responding schools – the variations in outcomes from this method was not desirable.

6) The availability of a variable (the 2008 Numeracy mean) that was very strongly correlated with the 2009 Numeracy mean, meant that the regression based approaches generally performed well, with minimal bias and good precision. This strong relationship also appeared to reduce the effects of NMAR bias for the regression based approaches.

7) The method of school substitution performs very well – the best of all the approaches - with respect to minimising bias and maintaining good precision. The use of the school-substitution method results in the maintenance of the sample size on which to base estimates of outcomes, even under non-response, and this contributes substantially to maintaining good precision with this approach.

At least with respect to this particular sub-population of Australia’s school system, the results provide confirmation that the currently used practice of school substitution is useful for managing non-response and should contribute to estimates of outcomes with less bias and better precision. When school non-response extends beyond that which can be managed by school-substitution, for example when neither the sampled school nor its designated replacements participate in a survey, adjustment approaches that can take account of prior performance and socio-economic background should be preferred using regression imputation or a regression estimator.
In the next chapter we consider constructing a simulated student population that will be used in Chapter 8 to examine how non-response adjustment methods perform when there is non-response at the student level.
Chapter 7  Constructing a simulated student population

7.1  Introduction

Chapter 6 investigated a range of methods to address non-response at the school level. As previously discussed, schools are essentially mandated to participate in surveys conducted under Australia’s NAP, and so very little school-level non-response is observed for these surveys. School level response rates for TIMSS 2011 were 96% of originally sampled schools, and 98% after including replacement schools (Table 4-2). With no other empirical evidence for the drivers of school non-response in a non-mandated environment, as experienced for surveys conducted outside the NAP, a starting point for the investigation into school-level non-response in the last chapter was
to link to the observed student-level response rates achieved in the TIMSS survey. These analyses (Chapter 5) showed that response rates at the student level varied across a number of population characteristics, for example by prior performance and by socio-economic background. If student participation varied across these characteristics, then schools catering for students with those characteristics might be similarly more or less likely to respond in a non-mandated environment.

On the other hand, the drivers for non-response at the school and student levels might be quite different. One example is that of larger schools from smaller jurisdictions. Because the two stage sample design used for most student surveys involves sampling schools with probability proportional to size, the larger schools from the smallest jurisdictions are almost always sampled for major surveys and so the burden of participation at the school level, which includes the disruption to the school curriculum program, the involvement of teaching staff in coordinating the activity, can fall disproportionately on these schools. While not necessarily completely independent of factors such as socio-economic background and prior performance, for example larger schools tend to be in more urban areas and location is correlated with socio-economic background, the set of factors leading to school-level non-response may be quite different to the factors leading to non-response at the student level.

Similarly, adjustments to address non-response at the school level may not have the same effectiveness at the student level. For example, the use of substitution proved to be an effective measure to address school level non-response for the population studied in Chapter 6. It was possible to find substitutes at the level of the institution that matched well with the sampled schools across the variables of interest. Substitution at the student
level however is generally not applicable, and is a practice that is not applied in most large scale student surveys. At this level, non-participation typically does not surface until the day of the assessment or shortly before. At that late stage, it is generally not practicable operationally to identify a well matched substitute. Participation in such surveys typically requires a formal process of informing parents of the survey and obtaining their consent, and these processes will not have been conducted for non-sampled students. In any case, while variations in individual student characteristics might reasonably be averaged over a group of students at a particular school, those individual variations mitigate against the task of identifying a well-matched student as a potential substitute for any particular sample student. It is important therefore to consider non-response at school and student levels separately.

The following chapter investigates approaches to managing non-response at the student level. To explore and evaluate models and methods for non-response adjustment at this level a population of students was required. As previously noted, the data collected through the NAPLAN activity provides Australian governments with comprehensive information about student progress, and such data would be an extremely valuable resource for managing non-response were it to be made available for this purpose. These data were not made available for the purposes of this research and so it was necessary to use available data to develop a realistic simulated student population to explore non-response management approaches under a range of student level non-response scenarios.

It was desirable that the simulated population reflects as closely as possible an actual population of students in the Australian context:
• from schools distributed by sector, location and socio-economic background that mirrored those distributions in a group of schools within Australia;

• with a similar profile of students with respect to gender, age range and ability as would appear in a typical set of schools from within that group.

This chapter provides a model for an approach to building a realistic simulated population of students making use of data from the (sample-based) TIMSS survey. This model might well prove useful for other countries that participated in TIMSS or other sample-based surveys to enable exploration of non-response effects and mitigation strategies at the student level within their own context.

The subpopulation of Victorian government schools that participated in the TIMSS 2011 survey was the starting point for developing a student population. It was possible to identify this subpopulation using the stratification information published for this survey (TIMSS (2013)). The explicit stratification used for Australia’s participation in TIMSS was State or Territory and the implicit stratification was a cross classification of location (metropolitan / provincial / remote) and sector (Catholic, government, independent). Drawing from the TIMSS survey establishes a link for this investigation with the earlier chapters explaining achievement and response in Australian educational surveys. Because it involves the same subpopulation, the investigation also links to Chapter 6 on school-level non-response.

The aim is to apply a range of non-response mitigation strategies at this level, such as weighting adjustment, response propensity models, regression based methods and imputation, under different assumptions concerning the non-response mechanism.
Missing completely at random (MCAR), missing at random conditional on covariates (MAR), and not missing at random (NMAR) are considered and the performance of the non-response adjustment strategies at the student level under the various conditions are evaluated. The role of auxiliary variables used in the adjustments is examined.

7.2 Constructing a student level population file

543 of the 6523 records on the Australian TIMSS student database for Year 8 are students from 24 Victorian government schools. The estimated population of students from these 24 schools was 4202. The aim was therefore to begin with the 543 records from the TIMSS database and add 3659 further records through imputation, making use of the 6523 records for students across Australia, so as to obtain a student level data file with 4202 records.

As discussed in Chapter 2, section 2.6.5, imputation is an approach of substituting a synthetic value for a real value, although the synthetic value may also be a real value from another hopefully similar unit in the data set. The discussion in Chapter 2 was in relation to substitutes for non-responding units, but in this chapter the imputations are for non-sampled students from the sampled schools that participated in TIMSS.

After exploring a range of different approaches to imputing the required records the decision was made to use a nearest-neighbour type algorithm as explained below. As the investigation was focussed on the effects and management of unit-level non-response where no information was obtained directly from non-participating students, it was important that the imputed cases authentically represented the complexity of data that occurs at the individual student level. For each non-sampled student from Victorian
government schools that were sampled for TIMSS, a full record of a student from another part of the Australian TIMSS dataset will be added. This was considered superior to more theoretical imputation methods, for example where imputations might be drawn from a regression model. By making these imputations, a simulated population file of all students enrolled at year 10 in a set of schools sampled for TIMSS will have been constructed.

7.3 Imputation algorithm

The steps below summarise the algorithm used for imputing the student records:

1. All schools from the TIMSS database, from all States and sectors, were grouped into five classes based on quintiles of the SEIFA Index of Education and Occupation (IEO) for the school. As noted in Chapter 4, this is a school-postcode based measure of socio-economic background, a variable that was noted to be related to outcomes on this assessment.

2. Within each SEIFA quintile the average mathematics outcome for TIMSS for the school was plotted against the NAPLAN Numeracy mean at Year 9 for the school. Both measures were scaled with a standard deviation of 100.

3. The Euclidean distance between these two measures was calculated between each of the Victorian government schools and all other schools in the TIMSS database, and schools whose distance was within half a standard deviation (50) were identified as ‘near neighbours’.

The following summarises the distance algorithm used to find neighbours for each school:

239
Let $x_i$ be the mean 2010 Year 9 NAPLAN maths score for school $i$

Let $y_i$ be the TIMSS Maths outcome for school $i$

Let School $a$ be the school for which imputations are required

For all other schools ($b$) that participated in TIMSS and are in the same SEIFA quintile: if \[ \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2} < 50 \] then school $b$ is considered ‘close’ to school $a$

The algorithm is illustrated in Figure 7-1 with the following example involving schools in the TIMSS database from the third SEIFA quintile. The Victorian schools for which records are required to be imputed are in a different colour to all other schools. The codes for these schools are 513 514 515 516 517 and 536.

Figure 7-1: TIMSS mean versus NAPLAN mean – Victorian government versus other jurisdictions
Taking, for example, one of the six schools from this quintile for which imputations are required, school 515, Figure 7-2 below highlights in blue the set of schools identified as neighbours according to the above algorithm.

Figure 7-2: Scatterplot of TIMSS mean versus NAPLAN mean – Near neighbours

The above analysis was run for the 24 schools requiring imputations to see whether enough data was available from neighbouring schools as defined in the above algorithm.
to successfully impute data. Table 7-1 in the following section summarises the number of records available.

One of the 24 schools was identified as an outlier in the analysis. It is the school ‘516’ that appears at the bottom of Figure 7-2 above, shown as quite apart from the other schools in this IEO quintile. Further investigation revealed that this school had radically transformed from one school to four separate schools around the period that the NAPLAN measure and the TIMSS measure was generated. On this basis it was decided that this school should be dropped from further analysis. For the remainder of the analysis, the simulated population would be built from 23 schools, with 517 records, to achieve a population size after imputation of 4176 students.

7.4 Records available for imputation

Table 7-1 summarises the number of student records identified from neighbouring schools for each of the schools requiring imputations. In all cases there were sufficient records available from neighbouring schools to build the population of students.
A further step was to separate the total enrolment for each school into the number of boys and number of girls and to make the imputations separately for boys and girls so that the distribution matched with the school distribution.

The simulated population was built by random sampling without replacement from the pool of students (girls and boys respectively) available from the neighbouring schools and appending these imputed records to the original records for the school.

### 7.5 Imputation outcomes at the school level

#### 7.5.1 Mathematics mean

Table 7-2 summarises the results of the imputation with respect to the mean TIMSS mathematics outcome. The number of cases as well as the mean and standard deviation of the original records, imputed records and for the original and imputed records
combined for each school are shown. Flags are identified for cases where the difference in the mean or the standard deviation between the combined records and the original records is more than half a standard deviation as estimated with the original records, conventionally considered a ‘medium’ effect size, when comparing two means, (Cohen, 1992).

For example in the case of school 503, the mean Mathematics score of the 320 cases after imputation was 573.5, about 6 points lower than the mean of the 22 records from this school who participated in TIMSS. The difference in mean scores is within one tenth of an overall standard deviation, which was considered acceptable. For school 531, the difference in means between the original records and the full records after imputation exceeded half a standard deviation. This is the one case where the criterion of half a standard deviation is not met. Overall, with respect to the mathematics performance profile of the schools, the imputations have produced a realistic distribution of schools.
Table 7-2: Comparison of original records and imputed records: TIMSS 2011 Science Year 8 Mathematics

<table>
<thead>
<tr>
<th>School</th>
<th>Original records</th>
<th>Imputed records</th>
<th>Total combined records</th>
<th>Difference flags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
<td>SD</td>
<td>n</td>
</tr>
<tr>
<td>503</td>
<td>22</td>
<td>579.8</td>
<td>66.8</td>
<td>298</td>
</tr>
<tr>
<td>504</td>
<td>26</td>
<td>538.2</td>
<td>71.2</td>
<td>178</td>
</tr>
<tr>
<td>505</td>
<td>25</td>
<td>657.3</td>
<td>49.4</td>
<td>126</td>
</tr>
<tr>
<td>506</td>
<td>18</td>
<td>479.1</td>
<td>90.9</td>
<td>52</td>
</tr>
<tr>
<td>507</td>
<td>24</td>
<td>532.6</td>
<td>60.9</td>
<td>178</td>
</tr>
<tr>
<td>508</td>
<td>23</td>
<td>517.3</td>
<td>40.4</td>
<td>247</td>
</tr>
<tr>
<td>509</td>
<td>23</td>
<td>543.4</td>
<td>45.9</td>
<td>218</td>
</tr>
<tr>
<td>510</td>
<td>22</td>
<td>504.9</td>
<td>60.1</td>
<td>178</td>
</tr>
<tr>
<td>511</td>
<td>22</td>
<td>439.7</td>
<td>68.3</td>
<td>169</td>
</tr>
<tr>
<td>513</td>
<td>23</td>
<td>472.6</td>
<td>55.0</td>
<td>95</td>
</tr>
<tr>
<td>514</td>
<td>20</td>
<td>522.3</td>
<td>53.0</td>
<td>176</td>
</tr>
<tr>
<td>515</td>
<td>21</td>
<td>519.3</td>
<td>74.9</td>
<td>264</td>
</tr>
<tr>
<td>517</td>
<td>23</td>
<td>453.1</td>
<td>63.3</td>
<td>115</td>
</tr>
<tr>
<td>518</td>
<td>46</td>
<td>466.9</td>
<td>68.0</td>
<td>215</td>
</tr>
<tr>
<td>519</td>
<td>14</td>
<td>475.1</td>
<td>56.5</td>
<td>130</td>
</tr>
<tr>
<td>520</td>
<td>19</td>
<td>406.9</td>
<td>44.2</td>
<td>51</td>
</tr>
<tr>
<td>521</td>
<td>24</td>
<td>488.8</td>
<td>73.3</td>
<td>135</td>
</tr>
<tr>
<td>531</td>
<td>20</td>
<td>387.4</td>
<td>38.3</td>
<td>127</td>
</tr>
<tr>
<td>532</td>
<td>12</td>
<td>433.0</td>
<td>52.9</td>
<td>209</td>
</tr>
<tr>
<td>533</td>
<td>26</td>
<td>478.8</td>
<td>67.9</td>
<td></td>
</tr>
<tr>
<td>534</td>
<td>23</td>
<td>462.8</td>
<td>59.9</td>
<td>136</td>
</tr>
<tr>
<td>535</td>
<td>26</td>
<td>466.8</td>
<td>61.2</td>
<td>185</td>
</tr>
<tr>
<td>536</td>
<td>18</td>
<td>496.0</td>
<td>64.1</td>
<td></td>
</tr>
<tr>
<td>537</td>
<td>23</td>
<td>465.8</td>
<td>66.8</td>
<td>177</td>
</tr>
</tbody>
</table>

Figure 7-3 plots the imputed school mean mathematics score with the mean from the original sample of students from the school. There is a very strong relationship between the two ($r = 0.99$).
Figure 7-3: TIMSS 2011 Y8 Mathematics mean scores – original versus combined records

Figure 7-4 plots the standard deviations of the mean scores for the combined records following imputation versus the original records. Overall the relationship between the two variables is notably weaker ($r = 0.38$). Many schools sit close to the regression line, but for a small number of schools, the variation in scores is greater across the combined records compared to the original records or vice versa. The aim of the imputation approach is to achieve a realistic degree of variation within schools and that has been achieved. The standard deviations of the original records vary more across schools than for the combined records, which is expected as the original records are based on a smaller sample.
1.4.1 Means for Science Achievement, Age and Home Educational Resources Index

Table 7-3 to Table 7-5 show the same comparisons for three other variables: the mean science achievement, the age of students and the mean score on the Home Educational Resources (HER) measure.
Table 7-3: Comparison of original records and imputed records: TIMSS 2011 Science Year 8

<table>
<thead>
<tr>
<th>School</th>
<th>Original records</th>
<th>Imputed records</th>
<th>Total combined records</th>
<th>Difference flags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
<td>SD</td>
<td>n</td>
</tr>
<tr>
<td>503</td>
<td>22</td>
<td>557.0</td>
<td>66.4</td>
<td>298</td>
</tr>
<tr>
<td>504</td>
<td>26</td>
<td>544.0</td>
<td>82.1</td>
<td>178</td>
</tr>
<tr>
<td>505</td>
<td>25</td>
<td>644.7</td>
<td>51.8</td>
<td>126</td>
</tr>
<tr>
<td>506</td>
<td>18</td>
<td>485.9</td>
<td>79.4</td>
<td>52</td>
</tr>
<tr>
<td>507</td>
<td>24</td>
<td>542.1</td>
<td>65.7</td>
<td>178</td>
</tr>
<tr>
<td>508</td>
<td>23</td>
<td>534.3</td>
<td>46.2</td>
<td>247</td>
</tr>
<tr>
<td>509</td>
<td>23</td>
<td>550.2</td>
<td>45.6</td>
<td>218</td>
</tr>
<tr>
<td>510</td>
<td>22</td>
<td>524.6</td>
<td>62.5</td>
<td>178</td>
</tr>
<tr>
<td>511</td>
<td>22</td>
<td>460.3</td>
<td>70.1</td>
<td>169</td>
</tr>
<tr>
<td>512</td>
<td>23</td>
<td>500.5</td>
<td>59.3</td>
<td>95</td>
</tr>
<tr>
<td>513</td>
<td>20</td>
<td>510.9</td>
<td>51.4</td>
<td>176</td>
</tr>
<tr>
<td>515</td>
<td>21</td>
<td>521.1</td>
<td>77.6</td>
<td>264</td>
</tr>
<tr>
<td>517</td>
<td>23</td>
<td>485.1</td>
<td>74.5</td>
<td>115</td>
</tr>
<tr>
<td>518</td>
<td>46</td>
<td>463.6</td>
<td>68.2</td>
<td>215</td>
</tr>
<tr>
<td>519</td>
<td>14</td>
<td>480.7</td>
<td>58.5</td>
<td>130</td>
</tr>
<tr>
<td>520</td>
<td>19</td>
<td>389.5</td>
<td>50.0</td>
<td>51</td>
</tr>
<tr>
<td>521</td>
<td>24</td>
<td>495.3</td>
<td>68.7</td>
<td>135</td>
</tr>
<tr>
<td>531</td>
<td>20</td>
<td>417.4</td>
<td>53.1</td>
<td>127</td>
</tr>
<tr>
<td>532</td>
<td>12</td>
<td>456.2</td>
<td>54.8</td>
<td>209</td>
</tr>
<tr>
<td>533</td>
<td>26</td>
<td>497.2</td>
<td>60.1</td>
<td>26</td>
</tr>
<tr>
<td>534</td>
<td>23</td>
<td>483.4</td>
<td>64.8</td>
<td>136</td>
</tr>
<tr>
<td>535</td>
<td>26</td>
<td>487.9</td>
<td>70.6</td>
<td>185</td>
</tr>
<tr>
<td>536</td>
<td>18</td>
<td>525.5</td>
<td>75.9</td>
<td>18</td>
</tr>
<tr>
<td>537</td>
<td>23</td>
<td>494.9</td>
<td>68.0</td>
<td>177</td>
</tr>
</tbody>
</table>
Table 7-4: Comparison of original records and imputed records: TIMSS 2011 Year 8:
Student Age

| School | Original records | | Imputed records | | Total combined records | | Difference flags |
|--------|------------------|--------|------------------|--------|------------------------|--------|
|        | n | Mean | SD | n | Mean | SD | n | Mean | SD | Mean | SD |
| 503    | 22 | 14.2 | 0.4 | 298 | 14.0 | 0.5 | 320 | 14.0 | 0.5 | 1   |
| 504    | 26 | 14.3 | 0.4 | 178 | 14.0 | 0.4 | 204 | 14.0 | 0.4 | 1   |
| 505    | 25 | 14.1 | 0.3 | 126 | 14.0 | 0.4 | 151 | 14.0 | 0.4 | 1   |
| 506    | 18 | 14.3 | 0.4 | 52  | 14.1 | 0.6 | 70  | 14.1 | 0.5 | 1   |
| 507    | 24 | 14.2 | 0.4 | 178 | 14.0 | 0.4 | 202 | 14.0 | 0.4 | 1   |
| 508    | 23 | 14.3 | 0.3 | 247 | 14.0 | 0.4 | 270 | 14.0 | 0.4 | 1   |
| 509    | 23 | 14.2 | 0.3 | 218 | 14.0 | 0.4 | 241 | 14.0 | 0.4 | 1   |
| 510    | 22 | 14.4 | 0.4 | 178 | 14.0 | 0.5 | 200 | 14.1 | 0.5 | 1   |
| 511    | 22 | 14.3 | 0.5 | 169 | 14.0 | 0.5 | 191 | 14.0 | 0.5 | 1   |
| 513    | 23 | 14.3 | 0.4 | 95  | 14.0 | 0.5 | 118 | 14.0 | 0.5 | 1   |
| 514    | 20 | 14.1 | 0.3 | 176 | 14.0 | 0.4 | 196 | 14.0 | 0.4 | 1   |
| 515    | 21 | 14.3 | 0.4 | 264 | 14.0 | 0.4 | 285 | 14.0 | 0.4 | 1   |
| 517    | 23 | 14.3 | 0.4 | 115 | 14.0 | 0.5 | 138 | 14.1 | 0.5 | 1   |
| 518    | 46 | 14.2 | 0.5 | 215 | 14.0 | 0.4 | 261 | 14.0 | 0.5 | 1   |
| 519    | 14 | 14.2 | 0.4 | 130 | 14.0 | 0.4 | 144 | 14.0 | 0.4 | 1   |
| 520    | 19 | 14.4 | 0.3 | 51  | 14.0 | 0.6 | 70  | 14.1 | 0.5 | 1   |
| 521    | 24 | 14.2 | 0.4 | 135 | 14.0 | 0.5 | 159 | 14.1 | 0.5 | 1   |
| 531    | 20 | 14.2 | 0.4 | 127 | 13.9 | 0.5 | 147 | 14.0 | 0.5 | 1   |
| 532    | 12 | 14.4 | 0.5 | 209 | 14.0 | 0.5 | 221 | 14.0 | 0.5 | 1   |
| 533    | 26 | 14.3 | 0.3 | 136 | 14.0 | 0.4 | 159 | 14.0 | 0.4 | 1   |
| 534    | 23 | 14.2 | 0.4 | 185 | 14.0 | 0.5 | 211 | 14.1 | 0.4 | 1   |
| 535    | 18 | 14.2 | 0.4 | 200 | 14.0 | 0.5 | 18  | 14.2 | 0.4 | 1   |
| 536    | 23 | 14.4 | 0.4 | 177 | 13.9 | 0.5 | 200 | 14.0 | 0.5 | 1   |
Table 7-5: Comparison of original records and imputed records: TIMSS 2011 Year 8
Home Educational Resources

<table>
<thead>
<tr>
<th>School</th>
<th>Original records</th>
<th>Imputed records</th>
<th>Total combined records</th>
<th>Difference flags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
<td>SD</td>
<td>n</td>
</tr>
<tr>
<td>503</td>
<td>22</td>
<td>11.5</td>
<td>1.4</td>
<td>298</td>
</tr>
<tr>
<td>504</td>
<td>26</td>
<td>11.8</td>
<td>1.6</td>
<td>178</td>
</tr>
<tr>
<td>505</td>
<td>25</td>
<td>12.2</td>
<td>1.4</td>
<td>126</td>
</tr>
<tr>
<td>506</td>
<td>18</td>
<td>11.1</td>
<td>1.6</td>
<td>52</td>
</tr>
<tr>
<td>507</td>
<td>24</td>
<td>11.5</td>
<td>1.6</td>
<td>178</td>
</tr>
<tr>
<td>508</td>
<td>23</td>
<td>12.3</td>
<td>1.3</td>
<td>247</td>
</tr>
<tr>
<td>509</td>
<td>23</td>
<td>11.6</td>
<td>1.5</td>
<td>218</td>
</tr>
<tr>
<td>510</td>
<td>22</td>
<td>11.7</td>
<td>1.6</td>
<td>178</td>
</tr>
<tr>
<td>511</td>
<td>22</td>
<td>10.5</td>
<td>1.3</td>
<td>169</td>
</tr>
<tr>
<td>513</td>
<td>23</td>
<td>11.1</td>
<td>1.7</td>
<td>95</td>
</tr>
<tr>
<td>514</td>
<td>20</td>
<td>10.6</td>
<td>1.3</td>
<td>176</td>
</tr>
<tr>
<td>515</td>
<td>21</td>
<td>10.6</td>
<td>1.0</td>
<td>264</td>
</tr>
<tr>
<td>517</td>
<td>23</td>
<td>11.0</td>
<td>1.2</td>
<td>115</td>
</tr>
<tr>
<td>518</td>
<td>46</td>
<td>9.8</td>
<td>1.0</td>
<td>215</td>
</tr>
<tr>
<td>519</td>
<td>14</td>
<td>10.7</td>
<td>1.2</td>
<td>130</td>
</tr>
<tr>
<td>520</td>
<td>19</td>
<td>9.3</td>
<td>1.3</td>
<td>51</td>
</tr>
<tr>
<td>521</td>
<td>24</td>
<td>9.4</td>
<td>1.6</td>
<td>135</td>
</tr>
<tr>
<td>531</td>
<td>20</td>
<td>10.0</td>
<td>1.3</td>
<td>127</td>
</tr>
<tr>
<td>532</td>
<td>12</td>
<td>10.3</td>
<td>1.3</td>
<td>209</td>
</tr>
<tr>
<td>533</td>
<td>26</td>
<td>11.2</td>
<td>1.8</td>
<td>26</td>
</tr>
<tr>
<td>534</td>
<td>23</td>
<td>10.8</td>
<td>1.2</td>
<td>136</td>
</tr>
<tr>
<td>535</td>
<td>26</td>
<td>10.1</td>
<td>1.5</td>
<td>185</td>
</tr>
<tr>
<td>536</td>
<td>18</td>
<td>11.1</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>537</td>
<td>23</td>
<td>11.1</td>
<td>1.6</td>
<td>177</td>
</tr>
</tbody>
</table>

Difference flags appear rarely on the mathematics and science measures. This is not surprising as the imputations were based on the mathematics outcome measure, and the correlation between the mathematics and science measure is high. Difference flags on the age and HER profile, neither of which were used in the imputation model, appear more often. However, an inspection of the sizes of the differences in each case suggests that imputations have been reasonably effective in producing a simulated dataset that broadly matches with typical Victorian school profile across a range of measures.
As with the mathematics imputation, the plot of the school science mean for the records after imputation with the school science mean based on the original records shows a very strong relationship ($r = 0.97$).

![Figure 7-5: TIMSS 2011 Y8 Science mean scores – original versus combined](image)

The relationship between the standard deviations of the original records and the combined records is a relatively weak one, as shown in Figure 7-6 ($r = 0.2$). Again the imputed data display a realistic degree of variation within schools as reflected in the school standard deviations.
As expected, the relationship between the HER of imputed and original records is less strong although the relationship is still quite strong ($r = 0.78$).

As with the mathematics and science comparisons, the relationship between the standard deviations of the original and the combined records after imputations is fairly weak ($r=0.26$).
7.5.2 Assigning students to classes

Because the sample design to be used on the simulated dataset involved the selection of an intact class of students from the sampled schools, it was important to take the further step of allocating students in the database into classes in a manner where the composition of students in those classes was as similar as possible to that observed in typical surveys of Australian students.

As a starting point, comparisons were made of the intra-class correlations on the key outcome measures for students in the original TIMSS database - with one class of student data from each school - with the intra-class correlations for the imputed dataset, with data for all students at the school. The intra-class correlation (ICC) is a measure of the similarity of students within classes. As the school ‘516’ had been dropped from the simulated population, it was also dropped in the analyses of the TIMSS data.
ICCs were calculated using M-Plus software Muthén and Muthén (2017), with sampling weights specified at the two stages of sample selection, schools then students, as proposed by Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash (1998).

Table 7-6 shows ICCs averaged over 100 separate samples – taken as the proportion of the total variance that is due to the between school variance in outcomes for each of three variables – HER, mathematics outcome and science outcome.

<table>
<thead>
<tr>
<th>Variable</th>
<th>TIMSS sample (intact class)</th>
<th>Imputed dataset (school level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>0.357</td>
<td>0.339</td>
</tr>
<tr>
<td>Science</td>
<td>0.309</td>
<td>0.275</td>
</tr>
<tr>
<td>Home Educational Resources</td>
<td>0.135</td>
<td>0.100</td>
</tr>
</tbody>
</table>

For each variable, the intra-class correlation for the combined dataset after imputation is lower than that observed using the TIMSS sample data. One expects a higher ICC for an individual class sample than would be observed across students from all classes at the school, especially in subjects such as mathematics and science where some level of clustering due to the influence of the teacher and/or class placement practices at the school level – for example students in different study programs - may occur. The higher degree of clustering for data involving intact classes has been observed in many surveys.
For example, in TIMSS sampling documentation it is suggested that if countries are seeking an estimate of their intra-class correlation (IC) from previous surveys … “if a national centre has values of the IC based on students sampled within schools, such as the 1991 survey of International Assessment of Education Progress (IAEP), or the OECD Programme for International Student Assessment (PISA), it would be prudent to add 0.2 to the value of the IC to estimate the IC for classrooms sampled within schools - the focus of the TIMSS sample design” P. J. Foy, Marc (2001).

As part of allocating students to classes, it was therefore desirable to factor into the allocation a small degree of additional clustering, so that the intra-class correlations were more similar to those observed in the original dataset is described. This was achieved as follows:

- Based on the enrolment at the year level, the number of classes to be assigned was determined. Class sizes would be approximately equal in size, with approximately 25 students assigned to each class.
- Prior to class allocation, all students within the school were ranked by their maths outcome.
- Every 4th student from this ranking was fixed by this ranking. This meant that for each class approximately 6 students (25/4) similarly ranked students would be assigned.
- The remaining students were then randomly sorted and allocated to classes.
- Students were allocated to classes such that all classes were allocated equally, with approximately 25 students from each class.

After allocation, 100 samples of one randomly selected class from every school were selected, and the ICC for each sample was calculated. The average ICC from those 100 samples was 0.433 with a standard deviation of 0.046. Table 7-7 shows the results of this allocation for mathematics, science and the home educational resources index. Especially
with respect to the maths and science measures, the intra-class correlations now resemble more closely those observed in the TIMSS data.

As the clustering algorithm was based on the mathematics outcome, the estimated ICC with respect to the less correlated home educational resources measure is further from that observed with the TIMSS data. Furthermore, as observed in the comparison of mean outcomes and standard deviations between original and imputed records, a larger difference with respect to the HER measure is also at least partially explained by the fact that this measure was not used in the model for identifying close schools.

Table 7-7: Intra-class correlations with additional clustering factor added to imputation model

<table>
<thead>
<tr>
<th>Variable</th>
<th>TIMSS sample (intact class)</th>
<th>Imputed dataset (school level)</th>
<th>Imputed dataset with a clustering factor in class allocation. (Average of 100 samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>0.357</td>
<td>0.339</td>
<td>0.357</td>
</tr>
<tr>
<td>Science</td>
<td>0.309</td>
<td>0.275</td>
<td>0.279</td>
</tr>
<tr>
<td>Home Educational Resources</td>
<td>0.135</td>
<td>0.10</td>
<td>0.085</td>
</tr>
</tbody>
</table>

Table 7-8 shows the allocation of students to the classes from each school following this approach.
7.6 Summary of imputation outcomes

The simulated population produced using the imputation of actual records taken from the TIMSS dataset is realistic. It displays a good degree of variation on the key survey variables, whilst substantially maintaining the properties of students that were present in the original sample data. The simulated population should provide a good basis for assessing the effectiveness of different non-response adjustments at the student level, which will be done in Chapter 8.
Chapter 8  Estimation of student outcomes under non-response

8.1 Introduction

This chapter investigates and compares the effects of various approaches to the management of student level non-response under different scenarios of non-response using the population data file simulated using the methods described in Chapter 7.
The student data file consisted of 4176 students from 23 schools, which is a sample of Victorian Government secondary schools that had been selected with probability proportional to size for the TIMSS 2011 survey, the group of students who participated in that survey and a set of imputed records to bring the number of records for each school up to the Year 8 enrolment size. There are 39053 students in the larger population of Victorian government secondary schools from which this sample of 23 schools was drawn. It was assumed that the population size and school enrolment data are known without error prior to sampling.

The students within each school in the simulated population were distributed into classes of approximately equal in size at around 25 students. As discussed in Chapter 7, the allocation of students to classes included an additional clustering factor so that intra-class correlations were of a similar order to those observed in the TIMSS survey data.

For this investigation of student level effects of non-response, the full participation of the schools sampled to participate in the survey was assumed so that variation in outcomes arising from non-response would be attributable to factors at the student level.

Section 8.2 provides an overview of the approach to the investigation. Section 8.3 details the design weights for unit records, reflecting the probabilities of selection of schools then students. Following a discussion in section 8.4 of the variables used in the analysis, section 8.5 describes the outcomes to be estimated and section 8.6 summarises regression modelling to identify the key variables that explain variation in these outcomes. Section 8.7 summarises the respective methods for inducing non-response in the data file under the non-response scenarios considered. Section 8.8 then summarises
the methods that were used to adjust the data to account for non-response. Sections 8.9 to 8.11 examine the effects of the adjustment methods under each of the non-response scenarios, with an overall summary of findings from the student level non-response investigation in section 8.12.

### 8.2 Approach to the investigation

One thousand student samples were randomly selected for this investigation. Each sample consisted of one class selected with equal probability from each of the 23 schools in the data file. The size of each sample was therefore quite consistent at approximately 568 students, made up of one school with 18 students and 22 schools each with approximately 25 students.

For each selected sample, non-response was induced according to various algorithms described in section 8.7 below. A range of adjustment methods (section 8.8) were then used to adjust for this non-response, and estimates of outcomes were derived.

There were some cases, particularly under scenarios with a high level of non-response, where the selected sample had properties that made them unsuitable for comparison across adjustment methods, or across the samples drawn for a particular adjustment method.

It was possible, for example, that a sample would be drawn with no responding students at all from a particular school. As this investigation was focussed on approaches to non-response within schools with an assumption of full school response, any sample with no
responding students from a school was removed from the investigation. This happened in a very small number of cases in the final model involving NMAR non-response.

Another issue was that in the calculation of the generalised regression estimator based on the prior performance, HER and sex of the student (adjustment method 6, section 8.8.7), the sum of these auxiliary variables by the non-response adjusted weight did not resolve to the respective population sums as expected. In other words, the generalised regression based adjustments calculated for some samples were not calibrated to the auxiliary information in the manner described in section 2.7.5. The issue did not occur at all under the mildest non-response scenarios (MCAR, MCAR2 or MAR1). It arose with the MAR2 model in 63 samples from 1000 drawn, for one sample in NMAR, about one in four samples under NMAR2 and about two in three samples under NMAR4. The issue appeared to be related to a combination of the levels of non-response and the complexity of weighting under the clustered sample design.

Where samples were considered unsuitable for comparison, the sample was removed and replaced with another sample. This approach ensured that comparisons across the adjustment methods were based on the same 1000 samples selected for each non-response scenario. It also ensured that within each adjustment approach, the distributions and summary statistics were all based on samples sharing the same basic characteristics.

For each of the 1000 samples selected for comparison under each non-response scenario, the average of the outcomes under each adjustment approach from the 1000 samples was taken as an empirical estimate of the expected value of the estimate taking
into account both the sampling and the non-response adjustments. The difference between the expected value and the population value was an indicator of bias. The standard deviation of outcomes from these 1000 samples – i.e. an empirical estimate of the standard error - served as an indicator of precision. The distribution of estimates under each adjustment are displayed through boxplots. Tables summarising the estimated mean square error and the root mean square error are also provided.

8.3 School and within-school design weights

The schools sampled for TIMSS were selected with probability proportional to size. The sampling interval for the school sample selection was therefore the population size divided by the number of sampled schools: 39053/23 = 1698. The school selection probability was the school measure of size – the enrolment size at the target grade - divided by the sampling interval, and the school design weight was the inverse of this school selection probability.

The selection within schools consists of the equal probability sampling of one intact class from each school. The within school selection probability is therefore \( \frac{1}{n(\text{classes})} \) and the within school design weight is its inverse, the number of classes.

For the investigations in this chapter, it is assumed that there is no school non-response, so every sample involves 23 classes of students from 23 schools. The effect of school-level non-response was examined in Chapter 6.
The final student design weight was the product of the school design weight and the within school design weight. Because the school weight and number of classes were constant for each school, the design weight for each student was invariant sample to sample. Table 8-1 summarises the selection probabilities and design weights for each school in the population.

Table 8-1: Design weights by school ID. Simulated population.

<table>
<thead>
<tr>
<th>School ID</th>
<th>Number of students</th>
<th>Number of classes at year 8</th>
<th>Sampling Interval</th>
<th>Selection probability</th>
<th>School design weight</th>
<th>Within school design weight</th>
<th>Overall student design weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>503</td>
<td>320</td>
<td>13</td>
<td>1698</td>
<td>0.19</td>
<td>5.31</td>
<td>13</td>
<td>68.98</td>
</tr>
<tr>
<td>504</td>
<td>204</td>
<td>8</td>
<td>1698</td>
<td>0.12</td>
<td>8.32</td>
<td>8</td>
<td>66.59</td>
</tr>
<tr>
<td>505</td>
<td>151</td>
<td>6</td>
<td>1698</td>
<td>0.09</td>
<td>11.24</td>
<td>6</td>
<td>67.47</td>
</tr>
<tr>
<td>506</td>
<td>70</td>
<td>3</td>
<td>1698</td>
<td>0.04</td>
<td>24.26</td>
<td>3</td>
<td>72.77</td>
</tr>
<tr>
<td>507</td>
<td>202</td>
<td>8</td>
<td>1698</td>
<td>0.12</td>
<td>8.41</td>
<td>8</td>
<td>67.25</td>
</tr>
<tr>
<td>508</td>
<td>270</td>
<td>11</td>
<td>1698</td>
<td>0.16</td>
<td>6.29</td>
<td>11</td>
<td>69.18</td>
</tr>
<tr>
<td>509</td>
<td>241</td>
<td>10</td>
<td>1698</td>
<td>0.14</td>
<td>7.05</td>
<td>10</td>
<td>70.45</td>
</tr>
<tr>
<td>510</td>
<td>200</td>
<td>8</td>
<td>1698</td>
<td>0.12</td>
<td>8.49</td>
<td>8</td>
<td>67.92</td>
</tr>
<tr>
<td>511</td>
<td>191</td>
<td>8</td>
<td>1698</td>
<td>0.11</td>
<td>8.89</td>
<td>8</td>
<td>71.12</td>
</tr>
<tr>
<td>513</td>
<td>118</td>
<td>5</td>
<td>1698</td>
<td>0.07</td>
<td>14.39</td>
<td>5</td>
<td>71.95</td>
</tr>
<tr>
<td>514</td>
<td>196</td>
<td>8</td>
<td>1698</td>
<td>0.12</td>
<td>8.66</td>
<td>8</td>
<td>69.30</td>
</tr>
<tr>
<td>515</td>
<td>285</td>
<td>11</td>
<td>1698</td>
<td>0.17</td>
<td>5.96</td>
<td>11</td>
<td>65.54</td>
</tr>
<tr>
<td>517</td>
<td>138</td>
<td>6</td>
<td>1698</td>
<td>0.08</td>
<td>12.30</td>
<td>6</td>
<td>73.82</td>
</tr>
<tr>
<td>518</td>
<td>261</td>
<td>10</td>
<td>1698</td>
<td>0.15</td>
<td>6.51</td>
<td>10</td>
<td>65.06</td>
</tr>
<tr>
<td>519</td>
<td>144</td>
<td>6</td>
<td>1698</td>
<td>0.08</td>
<td>11.79</td>
<td>6</td>
<td>70.75</td>
</tr>
<tr>
<td>520</td>
<td>70</td>
<td>3</td>
<td>1698</td>
<td>0.04</td>
<td>24.26</td>
<td>3</td>
<td>72.77</td>
</tr>
<tr>
<td>521</td>
<td>159</td>
<td>6</td>
<td>1698</td>
<td>0.09</td>
<td>10.68</td>
<td>6</td>
<td>64.07</td>
</tr>
<tr>
<td>531</td>
<td>147</td>
<td>6</td>
<td>1698</td>
<td>0.09</td>
<td>11.55</td>
<td>6</td>
<td>69.30</td>
</tr>
<tr>
<td>532</td>
<td>221</td>
<td>9</td>
<td>1698</td>
<td>0.13</td>
<td>7.68</td>
<td>9</td>
<td>69.15</td>
</tr>
<tr>
<td>534</td>
<td>159</td>
<td>6</td>
<td>1698</td>
<td>0.09</td>
<td>10.68</td>
<td>6</td>
<td>64.07</td>
</tr>
<tr>
<td>535</td>
<td>211</td>
<td>8</td>
<td>1698</td>
<td>0.12</td>
<td>8.05</td>
<td>8</td>
<td>64.38</td>
</tr>
<tr>
<td>536</td>
<td>18</td>
<td>1</td>
<td>1698</td>
<td>0.01</td>
<td>94.33</td>
<td>1</td>
<td>94.33</td>
</tr>
<tr>
<td>537</td>
<td>200</td>
<td>8</td>
<td>1698</td>
<td>0.12</td>
<td>8.49</td>
<td>8</td>
<td>67.92</td>
</tr>
</tbody>
</table>

The school with a single class of 18 students, school 536, has a larger design weight (94.33) than all other schools. The design weights for the other 22 schools have little variation, centred at 68.6 and ranging from 64.1 to 73.8.

8.4 Key variables used in the analysis

For each student in the dataset there existed two different cognitive outcome measures from the TIMSS assessment – an estimate of mathematics achievement and an estimate of science achievement. The correlation between these two measures of achievement
was 0.83. For the investigations in this chapter, the mathematics outcome was taken as the dependent variable while the science outcome was used in the role of a prior performance measure known and available for all students in the database. Based on the analyses in chapters 4 and 5 of the key variables explaining achievement and response, other potentially important explanatory variables included the student level Home Educational Resources (HER) Index, student sex, and the SEIFA index of Education and Occupation (IEO) based on the postal location of the school.

For the scenarios investigated, the impact of student non-response was that the mathematics outcome for the non-responding student was not known. However, the other variables noted above were known for all students, whether or not they responded. This was intended to mimic the situation that exists in Australian education, where prior performance measures and socio-economic background measures should be known for all students through their participation in NAPLAN, so that when a separate sample survey of students is selected (e.g. for TIMSS or one of the national surveys undertaken as part of Australia’s NAP) it is at least theoretically possible that such measures could be used in the management of non-response for the survey in question.

8.5 The population values to be estimated

The population under consideration was the simulated dataset, comprising all students from the sample of 23 schools that had been selected with probability proportional to size from the population of all Victorian government schools. The analyses to be made were with respect to the following outcomes:
• the mean mathematics outcome for all students and schools in the simulated dataset weighted by the school weight.

• The weighted proportion of students in the sampled schools achieving a minimum benchmark in mathematics – a mathematics score of at least 430.

These represent two main indicators on interest to users of educational survey results. Mean outcomes are very standard measures used for example in comparisons across jurisdictions. The proportion achieving a benchmark is an indicator of an important characteristic of the distribution of achievements. Other measures such as the standard deviation can also be considered.

Note that as the values to be estimated were the outcomes of a sample, they may vary from what would normally be considered the population parameters in this circumstance – the estimate of the mathematics mean and the proportion of students achieving a minimum benchmark of at least 430 on the mathematics scale - for all students in Victorian government schools. Nevertheless, the simulated dataset contains the mathematics mean and the relevant weights for all records and so the weighted mean and proportion achieving the benchmark from this dataset can be considered themselves as characteristics of a population. The population characteristics had the value 497.81 of 80.8% respectively. These are the population estimates that would be obtained from this sample of schools with full response at the student level. They would be estimated from samples drawn from the simulated dataset, along with estimates of the corresponding standard error.
8.6 Models explaining variation in outcomes

As with the investigations in earlier chapters, an analysis of (single level) multiple regression models using the available variables was undertaken to identify the most important variables for explaining achievement on the mathematics outcomes for the simulated population. The analyses were conducted using PROC SURVEYREG in SAS so that the clustering of students within schools was taken into account in the standard error estimates. The school ID variable was designated as the primary sampling unit (PSU). As all schools were included in all samples, no stratification was relevant for this population.

Table 8-2 shows the results of a regression of mathematics achievement explained by science achievement (BSSSCI01), home educational resources (BSBGHER) and sex (ITSEX), unweighted. Table 8-3 shows the same variables in an analysis weighted by the School design weight.

| Parameter    | Estimate | Standard Error | Denominator DF | t Value | Pr > |t| | Standardized Estimate |
|--------------|----------|----------------|----------------|---------|-------|-----------------------|
| Intercept    | 74.71    | 18.76          | 22             | 3.98    | 0.001 | 0.00                  |
| BSSSCI01     | 0.79     | 0.03           | 22             | 25.01   | <.0001| 0.81                  |
| BSBGHER      | 2.16     | 0.67           | 22             | 3.24    | 0.004 | 0.04                  |
| ITSEX        | -3.88    | 1.69           | 22             | -2.29   | 0.032 | -0.03                 |

Table 8-2: Regression parameters: Mathematics achievement explained by science achievement, home educational resources and sex, unweighted

R-Sq 0.68
R-Sq (Adj) 0.68
The results are very similar in both analyses, which is not surprising as the school weights are consistent across the data file. In the unweighted analysis, the regression coefficient of each variable is statistically significant in explaining achievement. Overall, 68% of the variation in the mathematics outcome is explained by these variables. In the weighted analysis, the science achievement measure and home educational resources are statistically significant variables, and the model explains 69% of the variation.

Other models were explored, incorporating additional factors such as the SEIFA Index of Education and Occupation (IEO) and location. These additional variables led to a slight increase in the explanatory power of the model, with the R-squared statistic rising to 0.71. For the analyses in this chapter it was decided to focus on the three variables noted above (science performance, home educational resources and sex). These variables in combination explained almost as much of the variation in mathematics achievement as other models, and as they were all student level variables they allowed for finer level adjustments for students within schools.

### 8.7 Non-response scenarios investigated

A number of different scenarios were considered:
1. Estimation with the full participation of sampled students.

2. Estimation when some sampled students were missing completely at random (MCAR). Under this scenario the expectation was that unbiased estimates would be achieved regardless of the non-response management strategy that was used. The estimated standard error of estimates would be larger than under the full response scenario due to the smaller responding sample size. Different adjustment methods may have larger standard errors that others, due to for example greater variation in weights.

3. Estimation when missing data arising from the non-participation of sampled students was missing at random (MAR), conditional on other survey variables. Under this scenario, management approaches that took into account those other variables should successfully produce unbiased estimates, with standard errors reflecting the relative rates of missingness within classes defined by those levels as well as the properties of the adjustment methods.

4. Estimation when missingness was related to the mathematics outcomes in ways that could not be fully explained by other survey variables. In this case, the missing data was not missing at random (NMAR). Under this scenario, non-response adjustment methods may reduce bias due to the use of the adjustment variables, but some bias is likely to remain. Management strategies under this scenario would be evaluated with respect to the degree of bias observed in estimated outcomes, as well as the size of standard errors around estimates.

The investigation will evaluate the benefits of adjustment methods with respect to removing biases arising from non-response, against their costs with respect to any loss of precision arising from the increased complexity of the adjustment. It may be the case that a more precise estimate that is somewhat biased arising from one adjustment
approach may be preferred over an adjustment that is less biased but carries a larger standard error due to the increased complexity of the method. This will be examined using the mean square error (MSE) of the estimates.

8.7.1 Inducing missing completely at random (MCAR) non-response

For each sample, one class was randomly sampled from each of the 23 schools in the population, and all students from that class were included in the sample. Each sampled student was allocated a random number between 0 and 1, and those with a random number above a certain response rate were identified as non-respondents.

Two versions of MCAR were explored. The first version used the relatively high response rate that was achieved nationally for the TIMSS survey at Year 8. As noted in section 5.1 the overall student participation rate after the use of replacement schools, was 88%, and this was the benchmark with which to assign response according to the algorithm above. Under this scenario, the estimated expected value of the achieved sample size across the 1000 samples drawn was 495, with a standard error of 7.6. The unweighted response rate was 88% with a standard deviation of 1.3%. Given the very minimal variation in weights as described in section 8.3 there was consistently almost no difference between weighted and unweighted response rates in the analyses undertaken below. This model will be referred to in the text as ‘MCAR1’.

The second version (‘MCAR2’) involved much lower response rates, which as noted in Section 6.1 does occur in surveys where participation is voluntary. A response rate of 35% was selected for this model. Below this response rate it was estimated using binomial probabilities that from the 1000 samples selected, some samples would be
dropped due to the non-response of all students from particular schools from the sample, which was undesirable.

Using the algorithm above to allocate students as respondents or non-respondents, the average sample size was 78. The average response rate over 1000 samples was 35% with a standard deviation of 2%. All 1000 samples had responding students from every school.

8.7.2 Inducing missing at random (MAR) non-response

The first model explored that involved response propensity related to other survey variables (‘MAR1’) was based on the pattern of response observed in TIMSS across measures of prior performance and socio-economic background. Section 5.4.3 examined overall student response rates across prior performance categories defined by the school NAPLAN mean, and (school level) quintiles of the SEIFA IEO measure. As a starting point to examine the effect of student level non-response under MAR, a concatenation of student level measures of prior performance and socio-economic background respectively – categories of the Science achievement score and quintiles of the TIMSS Home Educational Resources (HER) index - will be used, assuming the same response rates as had been observed in the TIMSS sample.
Figure 8-1 mirrors that of Figure 5-4 other than at some locations where no cases were observed in the corresponding cross classification of NAPLAN mean and IEO quintile in the TIMSS dataset. For example at the right end of the chart where the categories representing the highest decile of science performance and the lowest quintiles of HER (p90plus_q1 – p90plus_q3) are added to this chart, while the corresponding cross classification in TIMSS had no cases in these categories. In case these categories appeared in the science performance / HER cross classification for the simulated student file, the assumed rate for the next quintile (p90plus_q4) was assigned to those lower quintiles.

As with the MCAR process, for each selected sample, every sampled student from the 23 schools in the simulated student data file was assigned a random number. In this case, the response rate varied according to the values of the cross classification of science performance with the HER quintile. For example, in the left most value
representing the lowest decile of science performance and the lowest quintile of HER (‘p10_q1’), the assumed response rate was 0.78. If the assigned random number was at that value or below, a sampled student within that p10_q1 category was a respondent, otherwise he or she was identified as a non-respondent.

The expected value of the sample size under this model, estimated over 1000 samples was 492, corresponding to a response rate of 87%.

Similarly to the MCAR models, this first model of MAR was based on a high response rate and so it was desired to also evaluate non-response management approaches in an environment with lower response rates. A second MAR model (‘MAR2’) was developed which factored in a substantial additional factor of non-response based on the category of science performance. Beginning with the response rates for the first MAR model (MAR1), students in the highest quartile for science performance had the same response rates as for MAR1, but as the category of science performance reduced, the response rate dropped by a further 10% for each category, down to 50% for the lowest category. Instead of the rate of 0.78 for the p10_q1 group noted above, the rate for this group dropped by 50% to 0.39. Across all HER quintiles within the same performance category, the same reduction to the response rates used for MAR1 applied. The resulting response rates are shown in Figure 8-2.

The expected value of the sample size under this model was estimated as 365, with a response rate (weighted and unweighted) of 65%.
While such response rates are hypothetical, one might imagine a scenario where the survey was voluntary and undertaken over a lunch period, perhaps with an incentive for those who display a high level of performance. Response rates would likely drop off very rapidly for less able students, more or less regardless of their socio-economic background.

8.7.3 Inducing non-response not explained by the survey design variables (NMAR).

Three NMAR models were explored involving additional variation in response related to mathematics outcomes and therefore not able to be fully accounted for by other survey variables.

As with the approach to developing MAR models, the first model (‘NMAR1’) was based on the pattern of response observed in TIMSS across measures of prior
performance (i.e. science) and socio-economic background. Conditional on the response rates related to science performance and home educational resources, a further decline in response was induced related to categories of mathematics performance. The rate of decline increased for successive lower levels of mathematics performance. The response rate pattern used for this model is presented in Figure 8-3.

Each category of mathematics performance is presented as a line on Figure 8-3. Students in the top decile of mathematics performance ‘p90plus’, represented by the top line on the chart, have the same response rate pattern that was considered in the first MAR model. Students in the lowest ‘p10’ category, have the same pattern with respect to science and HER variables, but response rates have dropped by a further 70%. The rate of decline in response increases as the mathematics performance category drops.
The average sample size under this model was 431, and the expected value of the response rate estimated over 1000 samples was 77%.

The second NMAR model (‘NMAR2’) began with lower level response rates on the basis of categories of prior performance and HER, and with further declining response rates for lowering levels of mathematics performance. In the lowest levels of prior (science) performance, HER and mathematics performance, response rates dropped to around 21%. These improved for students in higher categories of mathematics performance, and for higher categories of HER and prior performance.

Figure 8-4: Assumed response rates: Student non-response MAR (v2) model

The average sample size under this model was 259, with an average response rate, weighted and unweighted, of 46%.
The third model (‘NMAR3’) began with the same pattern of response investigated in the
second MAR model, but with further declines in non-response as mathematics
performance dropped.

![Figure 8-5: Assumed response rates: Student non-response NMAR (v3) model](image)

For students in the highest decile of mathematics performance, response rates dropped
from 75% in the top science performance / HER category to 31% in the lowest.
Response rates for students in the lowest decile of mathematics performance, dropped
from 32% to 14% over that same range. For three of the 1000 samples selected, a
school had no responding students. These samples were dropped from further analysis,
see section 8.2. The average sample size (over 997 samples) for this model was 229
students, and the average response rate was 41%.
8.8 Adjustment methods

The following summarises the range of approaches to managing non-responses that were investigated across the respective models. The adjustment methods applied cover the various types of approaches described in chapter two, for example the weighting class and post-stratification approaches based on auxiliary variables described in section 2.6.1; the formation of response homogeneity groups (section 2.6.3); the use of generalised regression estimation (section 2.7.4); modelling of individual response propensities (section 2.10); and single and multiple regression imputation (section 2.6.6). The use of unit substitution, an approach that was successfully applied at the school level (section 6.9.7), was not applied at the student level, as this adjustment method is generally not practical at the level of the student for the reasons discussed in section 2.6.5.

8.8.1 No non-response adjustment

Under this scenario, no non-response adjustment was made. It was simply assumed that the non-respondents were a missing completely at random (MCAR) subset of the sampled students.

8.8.2 Method 1: A school level non-response adjustment weight equal to the number who were sampled divided by the number who responded

The first adjustment involved the simplest approach to managing non-response, where all sampled responding students within a school were weighted up by the same adjustment factor to represent the sampled students from the school. The underlying assumption was that the sampled non-respondents and respondents from the school were no different with respect to the survey outcomes. It took no account of the sex of
the non-responding students, their socio-economic background or any other background information.

This adjustment simply adds a constant weight adjustment to sampled participating students from the school, all of whom begin with the same design weight. With respect to the relative contribution of individual students to estimates at the school level, it is in fact equivalent to making no non-response adjustment at all. The relative contribution to estimates of students from a school remain the same whether weighted or unweighted. Across schools, the variation in weights arising from varying degrees of non-response within schools will have the effect of lowering overall precision on the survey estimates.

8.8.3 Method 2: non-response adjustment by sex within school, assuming at least five responding boys and girls. Otherwise method 1.

This method was the same as the first, except that the sampled students within each school were divided into two groups on the basis of their sex, and a separate adjustment was made for each group. In order to minimise weight variation, this method required a minimum of five students within each sex group in order to apply the non-response adjustment by sex. Otherwise, all students were given the same non-response adjustment (method 1).
8.8.4 Method 3: Post-stratification adjustments to design weights by n(samp)/n(resp) within classes defined across the population by prior performance home educational resources and sex.

In method 3, non-response adjustments were made to the design weights through the formation of 32 weight classes across the responding sample: 4 science performance quartiles * 4 HER quartiles * 2 groups by sex. The adjustment weighted the responding students from a weight class to represent the sampled (responding and non-responding) students from that class.

Even when taken across the respondents from all schools, the number of students in some groups could be very small - fewer than 5 students. For example, there were very small numbers of students in the highest quartile on the prior performance measure who were in the lowest HER quartile. Once again, if there were fewer than 5 students in a weight class, collapsing of classes was undertaken – e.g. by combining the bottom two HER quartiles. As is the case with this type approach in practice, decisions about the collapsing of cells were somewhat arbitrary – a factor to be kept in mind when evaluating this method.

8.8.5 Method 4: estimating the response propensity using logistic regression, and adjusting the design weight by the inverse of the estimated propensity.

In this approach, the outcome of responding to the survey was predicted using a logistic regression model using the respondent data with the science outcome, sex and the Home Educational Resources (HER) index as predictors. The outcome of this analysis was an estimated response probability of response for each case. The design weights were then adjusted by the inverse of this response probability.
8.8.6 Method 5: non-response adjustments to the design weights within weighting classes defined by quintiles of the response propensities estimated in method 4.

Non-response adjustments were made to the design weights through the formation of five weight classes defined as quintiles of the response propensities calculated in model 8. The adjustment weighted the responding students from a class to represent the sampled (responding and non-responding) students from that class.

8.8.7 Method 6: Using a generalised regression estimator based on the relationship between maths score and prior performance, home educational resources with estimates weighted by the design weight.

In this approach, a generalised regression estimator was used. The respondent data was used to generate a regression model where the mathematics outcome was predicted by the science measure, the home educational resources measure and sex. The vector of regression coefficients estimated from this model was then used to adjust the mean estimate for mathematics obtained from respondents in the manner described in section 2.6.3. The student design weight was used in the estimate of the mean outcome from the responding sample, as well as in the estimates of regression coefficients applied to the auxiliary data. As noted in section 8.2 a key characteristic of the 1000 samples used in the investigation was that the sums of the auxiliary variables used in the generalised regression estimated, when weighted by the product of the design weight and the greg weight calculated via this adjustment summed to their respective population values. The same 1000 samples were used across all of the adjustment methods described in this section.
8.8.8 Method 7: An imputation of the maths outcomes using a regression model based to predict mathematics achievement by science performance, HER and sex, on the respondent data for each sample.

In this approach, the missing data on maths achievement as a result of student non-response were imputed through the construction of a linear regression based on the respondent data from each sample, where the maths outcome was predicted by science performance, home educational resources and sex.

Regression analyses were run both unweighted and applying the school weights, with very similar outcomes. The outcomes presented were derived from a regression model using the product of the design weights at the school and student levels as the weight for the analyses.

Using the above regression model, the imputed values of mathematics achievement for the non-responding students were calculated and added to the student data file. A weighted estimate of the overall mathematics mean for each sample was then obtained.

8.8.9 Method 8: multiple imputation of mathematics outcomes on prior performance, home educational resources and sex.

For each sample, a multiple imputation on mathematics outcomes, prior performance, home educational resources and sex was conducted using Proc MI in SAS. As the missing data was limited to one variable - the mathematics outcome - there was no monotone missing data patterns to consider and the default option for Proc MI of Markov Chain Monte Carlo (MCMC) method was applied. This assumes that the data are multivariate normally distributed and that the missing data are missing at random. By default the method produces five imputations of the missing mathematics score. For
each imputed value, the mean of the 1000 samples was estimated to generate a population mean estimate per imputed value. The final estimate for the population mean was taken as the average of these five population mean estimates.

8.9 Approach to presenting outcomes

8.9.1 Overall approach

The eight non-response adjustment methods described in the preceding section (8.8.2 to 8.8.9) were tested across seven non-response models (2 * MCAR, 2 * MAR, 3 * NMAR), for estimating two population characteristics – the mean and the proportion over a score of 430, the minimum benchmark in mathematics (section 8.5).

A selection of these investigations are presented in graphical and tabular displays below to illustrate the general findings. Following the presentation of selected results, summary charts and tables provide distributional and quantitative comparisons across the complete set of investigations in terms of bias, mean square error and relative mean square error in section 8.11.

A discussion summarising the overall findings of the student level non-response investigations follows these summary tables in section 8.12.

8.9.2 Graphical and tabular representation of selected outcomes

For the selection of investigations presented below results are presented as follows:

- Box plots that compare the distribution of estimates of outcomes (the estimated means and proportions) for each management approach across the 1000 samples. The distribution of outcomes under full response is presented as the first plot in
each presentation, which serves as a benchmark for comparing the non-response management approaches implemented.

- Tables comparing the estimated mean and standard error of outcomes for each approach, calculated as the average outcome across the 1000 samples and the standard deviation of outcome estimates. Once again, the outcomes under full response appear in the first row.

- For the methods that involved explicit manipulation of weights, some statistics were generated on the variations in weights to assist with evaluating the methods used:
  
  o The first statistic calculated was the ratio $\frac{\text{sum}(\text{weight}^2)}{\text{sum}^2(\text{weight})}$, which is known as the weighting effect (WE) (Dorofeev and Grant (2006)).
  
  o Dividing the actual sample size by WE produces the ‘calibrated sample size’ ($n_c$). It can be loosely referred to as the effective sample size of the sample having taken into account the weighting. “The word ‘calibrated’ was also chosen to emphasise that what we get is the result of calibration of our sample against known values or more authoritative estimates to make it more representative.” (Dorofeev and Grant (2006)). The precision of estimates can be estimated as that which would be achieved with the calibrated sample size under an equally weighted sample.
  
  o The coefficient of variation in the weights (CV$_{wt}$) - the standard deviation of the weights divided by the mean - was also calculated. This gives a comparable measure of the variation in weights across different designs.
The coefficient of variation in the weights and the weighting effect are related as follows:

\[ WE = 1 + CV^2_{wt} \]

- The mean of each statistic across the 1000 samples was calculated and reported as the expected values.

The weight diagnostics under each scenario and method were applied across both the estimates for the population mean and the estimate of the proportion of scores above 430.

### 8.10 Selected results by scenario and method

#### 8.10.1 Estimating the mathematics mean - MCAR1

Table 8-4 compares the estimates of means derived from 1000 samples with 88% non-response MCAR across the non-response adjustments (see section 8.7.1). The population mean is 497.81.

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>497.88</td>
<td>3.96</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>497.88</td>
<td>4.21</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>497.87</td>
<td>4.11</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>497.89</td>
<td>4.06</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>497.89</td>
<td>4.03</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>497.89</td>
<td>4.05</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>497.79</td>
<td>2.07</td>
</tr>
<tr>
<td>h</td>
<td>Regression imputation - science, HER, sex as predictors</td>
<td>497.87</td>
<td>4.03</td>
</tr>
<tr>
<td>i</td>
<td>Multiple regression - science, HER, sex as predictors</td>
<td>497.89</td>
<td>4.05</td>
</tr>
</tbody>
</table>

Figure 8-6 charts the distribution of mean scores across the 1000 samples for each method.
All methods produced unbiased outcomes, with expected values on, or very close to the population mean. The estimates based on the regression estimator are noticeably more precise than for other methods, even compared to the standard estimator under full response. For the other methods, as expected there is a slight inflation in the standard error due to the reduction in the respondent sample size. The standard errors for the methods involving weighting class adjustments are slightly more inflated than the regression based estimators, or the imputation methods. Leaving aside the issues of non-response management, it is clear that with good auxiliary data, regression-based estimators produce good gains in precision compared to standard estimators. This is a well-known property of the generalised regression estimator, Särndal et al. (1992).
8.10.2 Estimating the proportion with scores over 430 in Maths – MCAR1

Table 8-5 compares the estimates of proportions of students scoring over 430 in mathematics across non-response adjustments under MCAR version 1. The population proportion is 0.808.

Table 8-5: Estimated proportions by non-response adjustment MCAR1

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>proportion</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>0.808</td>
<td>0.019</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>0.808</td>
<td>0.020</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>0.808</td>
<td>0.020</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>0.808</td>
<td>0.020</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>0.808</td>
<td>0.020</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>0.808</td>
<td>0.020</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>0.808</td>
<td>0.015</td>
</tr>
<tr>
<td>h</td>
<td>Regression imputation - science, HER, sex as predictors</td>
<td>0.813</td>
<td>0.020</td>
</tr>
<tr>
<td>i</td>
<td>Multiple regression - science, HER, sex as predictors</td>
<td>0.808</td>
<td>0.019</td>
</tr>
</tbody>
</table>

Figure 8-7 charts the distribution of proportions across the 1000 samples for each method.

Figure 8-7: Distribution of proportion estimates by non-response adjustment - MCAR1
Once again very few differences appear between the methods in their capacity to estimate the population proportion under MCAR non-response. The regression based imputation, method h, appears slightly biased compared to the other methods. As with the estimated means, the generalised regression estimator stands out as more precise compared to the other methods. The estimated standard errors are very similar across the other methods, with a slight inflation arising from the smaller sample size.

**Weight diagnostics - MCARI**

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Weight effect</th>
<th>Coefficient of Variation</th>
<th>Calibrated sample size</th>
<th>SE(Cal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>1.01</td>
<td>0.08</td>
<td>553.40</td>
<td>3.01</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>1.01</td>
<td>0.08</td>
<td>486.73</td>
<td>7.91</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within</td>
<td>1.02</td>
<td>0.13</td>
<td>480.96</td>
<td>8.72</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>1.01</td>
<td>0.09</td>
<td>485.36</td>
<td>8.12</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>1.01</td>
<td>0.09</td>
<td>485.94</td>
<td>8.01</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>1.01</td>
<td>0.09</td>
<td>485.79</td>
<td>8.03</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>1.01</td>
<td>0.11</td>
<td>483.58</td>
<td>8.26</td>
</tr>
</tbody>
</table>

The effect of the variation in weights on the precision of estimates is minimal across all methods under this scenario. The drop in the calibrated sample size for the adjustment methods compared to the sample with no non-response is almost entirely explained by the drop in the full sample size from around 568 (section 8.2) to around 500 due to the assumed rate of non-response.

**8.10.3 Estimating the mathematics mean - MCAR2**

Table 8-7 compares the estimates of means derived from 1000 samples with a much lower rate of response (35% compared to 88%) but where that non-response was MCAR (see section 8.7.1). The population mean is 497.81.
Even with much lower response rates, the methods still produce expected values at or very close to the population value. The regression estimator, method g, once again stands out as more precise than the other adjustment approaches. The standard errors have increased by around 25%, which is mostly explained by the reduction in sample size. However, as shown in Table 8-8 below, some of the weight adjustment methods show higher variation in weights in this scenario and this is having some effect on the
precision of estimates. The coefficient of variation in the weights is sometimes double or more the equivalent values in the MCAR1 scenario (Table 8-6).

**Weight diagnostics: MCAR2**

Table 8-8: Weight diagnostics: Estimates of population mean - MCAR2

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Weight effect</th>
<th>Coefficient of Variation</th>
<th>Calibrated sample size</th>
<th>SE(Cal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>1.01</td>
<td>0.08</td>
<td>553.59</td>
<td>3.02</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>1.01</td>
<td>0.08</td>
<td>194.16</td>
<td>11.20</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within</td>
<td>1.11</td>
<td>0.32</td>
<td>177.07</td>
<td>13.53</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>1.05</td>
<td>0.22</td>
<td>186.35</td>
<td>12.33</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>1.03</td>
<td>0.16</td>
<td>190.19</td>
<td>11.68</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>1.03</td>
<td>0.18</td>
<td>189.08</td>
<td>11.73</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>1.02</td>
<td>0.14</td>
<td>191.13</td>
<td>11.53</td>
</tr>
</tbody>
</table>

8.10.4 Estimating the proportion with scores over 430 in Maths – MAR1

Table 8-9 compares outcomes across the adjustment methods under a model with missing data similar to that observed in the TIMSS survey – high overall response rates but some missing data related to auxiliary variables (section 8.7.2).

Table 8-9: Estimated proportions by non-response adjustment MAR1

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>proportion</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>0.808</td>
<td>0.020</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>0.817</td>
<td>0.020</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>0.814</td>
<td>0.020</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>0.810</td>
<td>0.020</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>0.808</td>
<td>0.020</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>0.809</td>
<td>0.021</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>0.809</td>
<td>0.016</td>
</tr>
<tr>
<td>h</td>
<td>Regression imputation - science, HER, sex as predictors</td>
<td>0.812</td>
<td>0.020</td>
</tr>
<tr>
<td>i</td>
<td>Multiple regression - science, HER, sex as predictors</td>
<td>0.807</td>
<td>0.020</td>
</tr>
</tbody>
</table>

All methods produce outcomes close to the population proportion (0.808).
The boxplots comparison (Figure 8-9) indicates that the simple weighting class adjustments are slightly more biased than the other methods. The regression estimator approach (g) once again shows the best precision. All methods produce symmetric distributions with means and medians coinciding and a relatively small number of sample estimates some distance from the population value in both directions.

Figure 8-9: Distribution of proportion estimates by non-response adjustment – MAR1

**Weight diagnostics: MAR1**

Table 8-10 shows the weight diagnostics for the adjustments involving manipulation of weights under the MAR1 non-response scenario. The coefficient of variation for the weighting class adjustment by sex within school is higher, indicating some instability in the weights assigned to students under this approach. The calibrated sample sizes are similar across all methods under this non-response scenario.
8.10.5 Estimating the mathematics mean – MAR2

Table 8-11 compares outcomes across the adjustment methods under a model with higher levels of missing data but where that missing data is related to auxiliary variables, (section 8.7.2).

Table 8-11: Mean estimates by non-response adjustment. MAR2

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>497.52</td>
<td>3.88</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: ( n(\text{samp})/n(\text{resp}) ) by school</td>
<td>511.45</td>
<td>4.46</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: ( n(\text{samp})/n(\text{resp}) ) by sex within school</td>
<td>505.29</td>
<td>4.21</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>498.42</td>
<td>4.24</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>497.01</td>
<td>4.24</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>498.55</td>
<td>4.21</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>497.40</td>
<td>2.40</td>
</tr>
<tr>
<td>h</td>
<td>Regression imputation - science, HER, sex as predictors</td>
<td>497.09</td>
<td>4.24</td>
</tr>
<tr>
<td>i</td>
<td>Multiple regression - science, HER, sex as predictors</td>
<td>497.13</td>
<td>4.33</td>
</tr>
</tbody>
</table>

Outcomes from the weighting class adjustments that have not taken into account the factors relating to non-response are substantially biased. Unlike in the school non-response adjustment investigations, the distributions, while biased, are quiet symmetric, indicating that the adjustment methods are more stable with relatively more data points available in the case of student level non-response. The post-stratification adjustment that incorporates the prior measure of performance, home educational resources and sex
– substantially removes that bias. The response propensity approach has performed well, but the regression and imputation approaches, particularly the regression estimator, have been most successful in removing bias and keeping standard errors low.

![Figure 8-10: Distribution of mean estimates by non-response adjustment MAR2](image)

8.10.6 Estimating the mathematics mean – NMAR1

Table 8-12 compares outcomes across the adjustment methods with non-response related to outcomes and not able to be completely addressed with auxiliary variables available from the survey, (section 8.7.3).
Table 8-12: Mean estimates by non-response adjustment. NMAR1

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>497.86</td>
<td>3.97</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>507.11</td>
<td>4.17</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>503.43</td>
<td>4.05</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>500.89</td>
<td>4.10</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>500.13</td>
<td>4.10</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>500.82</td>
<td>4.14</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>500.04</td>
<td>2.24</td>
</tr>
<tr>
<td>h</td>
<td>Regression imputation - science, HER, sex as predictors</td>
<td>499.98</td>
<td>4.10</td>
</tr>
<tr>
<td>i</td>
<td>Multiple regression - science, HER, sex as predictors</td>
<td>500.01</td>
<td>4.13</td>
</tr>
</tbody>
</table>

None of the methods has completely removed the bias, but the methods that have taken into account prior performance and home educational resources have performed considerably better than the simpler methods that have not taken these factors into account.

The regression estimator once again stands out as having a much lower standard error.

With relatively high overall response rates under this scenario, the standard errors under
most of the other adjustments are only slightly inflated compared to those achieved under the full sample estimates.

8.10.7 Proportion with scores over 430 in Maths - NMAR1

A similar pattern of outcomes is observed for the adjustments used in the estimate of the population proportion. All methods are biased, but those that take account of factors contributing to non-response have performed substantially better than those that have not.

Table 8-13: Proportion estimates by non-response adjustment. NMAR1

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>proportion</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>0.808</td>
<td>0.019</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>0.846</td>
<td>0.019</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>0.834</td>
<td>0.020</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>0.824</td>
<td>0.021</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>0.820</td>
<td>0.021</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>0.823</td>
<td>0.021</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>0.825</td>
<td>0.017</td>
</tr>
<tr>
<td>h</td>
<td>Regression imputation - science, HER, sex as predictors</td>
<td>0.827</td>
<td>0.020</td>
</tr>
<tr>
<td>i</td>
<td>Multiple regression - science, HER, sex as predictors</td>
<td>0.818</td>
<td>0.020</td>
</tr>
</tbody>
</table>

The distributions of estimates of the proportion appear very slightly less symmetric with a wider range of more extreme observations at the top end of the distribution for a number of the adjustment methods employed.
8.10.8 Weight diagnostics: NMAR1

Table 8-14: Weight diagnostics: Estimates of population mean - NMAR 1

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Weight effect</th>
<th>Coefficient of Variation</th>
<th>Calibrated sample size</th>
<th>SE(Cal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>1.01</td>
<td>0.08</td>
<td>553.51</td>
<td>2.92</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>1.01</td>
<td>0.08</td>
<td>433.53</td>
<td>10.35</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>1.04</td>
<td>0.19</td>
<td>420.24</td>
<td>11.76</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>1.02</td>
<td>0.16</td>
<td>425.52</td>
<td>11.55</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>1.03</td>
<td>0.16</td>
<td>424.90</td>
<td>11.88</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>1.02</td>
<td>0.15</td>
<td>426.16</td>
<td>11.51</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>1.03</td>
<td>0.16</td>
<td>424.74</td>
<td>11.35</td>
</tr>
</tbody>
</table>

The weight variation is lower for the simplest weighting class adjustment method, as also indicated with a slightly larger calibrated sample size. While the weight variation is larger for the more complex methods, these have been more successful in reducing biased estimates.
8.10.9 Estimating the mathematics mean – NMAR3

Table 8-15 compares adjustment approaches with higher rates of NMAR3 non-response and a more complex relationship between response related to auxiliary variables NMAR non-response, (section 8.7.3).

Table 8-15: Mean estimates by non-response management method. NMAR3

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>495.57</td>
<td>3.73</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>522.23</td>
<td>4.45</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>511.42</td>
<td>4.37</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>502.83</td>
<td>4.63</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>500.44</td>
<td>4.85</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>502.63</td>
<td>4.64</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>502.07</td>
<td>3.27</td>
</tr>
<tr>
<td>h</td>
<td>Regression imputation - science, HER, sex as predictors</td>
<td>500.18</td>
<td>4.67</td>
</tr>
<tr>
<td>i</td>
<td>Multiple regression - science, HER, sex as predictors</td>
<td>500.22</td>
<td>4.78</td>
</tr>
</tbody>
</table>

Once again, no method has removed all of the bias, but the methods that do not take into account the MAR component of the non-response are considerably more biased compared to the other methods. Post-stratifying by prior performance, home educational resources and sex goes some way to reducing that bias, but further improvements have been made with the regression estimator methods, and the imputation methods. The standard errors are smaller for the regression estimator, and otherwise quite consistent across the regression estimator approaches, lower for these methods compared to most other adjustments.
The imputation based approaches, methods h and i, and the response propensity estimation, appear slightly less biased than the regression estimator, which has a smaller variation in estimates.

8.10.10 Weight diagnostics: NMAR3

Table 8-16 compares the weight diagnostics across the adjustment methods involving manipulation of weights. While the coefficient of variation in the weights is higher for the response propensity and regression based adjustments, these adjustments have resulted in less biased estimates, so have been worthwhile.

Table 8-16: Weight diagnostics: Estimates of population mean – NMAR3

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Weight effect</th>
<th>Coefficient of Variation</th>
<th>Calibrated sample size</th>
<th>SE(Cal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Full sample</td>
<td>1.01</td>
<td>0.08</td>
<td>553.69</td>
<td>3.00</td>
</tr>
<tr>
<td>b</td>
<td>Weight classes: n(samp)/n(resp) by school</td>
<td>1.01</td>
<td>0.08</td>
<td>232.80</td>
<td>12.29</td>
</tr>
<tr>
<td>c</td>
<td>Weight classes: n(samp)/n(resp) by sex within school</td>
<td>1.16</td>
<td>0.39</td>
<td>203.06</td>
<td>16.56</td>
</tr>
<tr>
<td>d</td>
<td>Post-stratification adjustment (science, HER, sex)</td>
<td>1.18</td>
<td>0.42</td>
<td>199.39</td>
<td>15.41</td>
</tr>
<tr>
<td>e</td>
<td>Response propensity adjustment</td>
<td>1.20</td>
<td>0.44</td>
<td>195.98</td>
<td>18.80</td>
</tr>
<tr>
<td>f</td>
<td>Post-stratification by response propensity classes</td>
<td>1.17</td>
<td>0.40</td>
<td>201.44</td>
<td>15.61</td>
</tr>
<tr>
<td>g</td>
<td>Regression estimator - design weights</td>
<td>1.12</td>
<td>0.35</td>
<td>208.61</td>
<td>12.21</td>
</tr>
</tbody>
</table>
8.11 Results summaries

8.11.1 Shape, precision and bias: Estimating the population mean - all scenarios

Figure 8-14 shows the observed distributions for each non-response management approach under each non-response model when estimating the population mean. Following those charts, Table 8-17 provide precision and bias statistics for each scenario.
Figure 8-14: Summary distributions – estimating the population mean – student level non-response
<table>
<thead>
<tr>
<th></th>
<th>EV</th>
<th>SE</th>
<th>bias</th>
<th>MSE</th>
<th>RMSE</th>
<th>EV</th>
<th>SE</th>
<th>bias</th>
<th>MSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>497.88</td>
<td>3.96</td>
<td>0.07</td>
<td>15.65</td>
<td>3.96</td>
<td>497.88</td>
<td>3.96</td>
<td>0.07</td>
<td>15.65</td>
<td>3.96</td>
</tr>
<tr>
<td>b</td>
<td>500.03</td>
<td>4.23</td>
<td>2.23</td>
<td>22.87</td>
<td>4.78</td>
<td>511.45</td>
<td>4.46</td>
<td>13.65</td>
<td>206.15</td>
<td>14.36</td>
</tr>
<tr>
<td>c</td>
<td>499.05</td>
<td>4.14</td>
<td>1.24</td>
<td>18.67</td>
<td>4.32</td>
<td>505.29</td>
<td>4.21</td>
<td>7.49</td>
<td>73.78</td>
<td>8.59</td>
</tr>
<tr>
<td>d</td>
<td>497.90</td>
<td>4.14</td>
<td>0.10</td>
<td>17.15</td>
<td>4.14</td>
<td>498.42</td>
<td>4.24</td>
<td>0.62</td>
<td>18.35</td>
<td>4.28</td>
</tr>
<tr>
<td>e</td>
<td>497.71</td>
<td>4.12</td>
<td>-0.09</td>
<td>16.94</td>
<td>4.12</td>
<td>497.91</td>
<td>4.24</td>
<td>-0.79</td>
<td>18.59</td>
<td>4.31</td>
</tr>
<tr>
<td>f</td>
<td>500.03</td>
<td>4.14</td>
<td>0.22</td>
<td>17.15</td>
<td>4.14</td>
<td>498.55</td>
<td>4.21</td>
<td>0.75</td>
<td>18.30</td>
<td>4.28</td>
</tr>
<tr>
<td>g</td>
<td>497.72</td>
<td>2.09</td>
<td>-0.09</td>
<td>4.36</td>
<td>2.09</td>
<td>497.40</td>
<td>2.40</td>
<td>-0.40</td>
<td>5.93</td>
<td>2.43</td>
</tr>
<tr>
<td>h</td>
<td>497.67</td>
<td>4.12</td>
<td>-0.14</td>
<td>16.98</td>
<td>4.12</td>
<td>497.09</td>
<td>4.24</td>
<td>-0.71</td>
<td>18.45</td>
<td>4.30</td>
</tr>
<tr>
<td>i</td>
<td>497.67</td>
<td>4.13</td>
<td>-0.13</td>
<td>17.08</td>
<td>4.13</td>
<td>497.13</td>
<td>4.33</td>
<td>-0.67</td>
<td>19.15</td>
<td>4.38</td>
</tr>
</tbody>
</table>

### Table 8-17: Precision and bias statistics – estimated mean under student level non-response
The comparison of distribution charts and summaries for the two MCAR models when estimating the population mean show that the major effect on lower response rates when data are missing completely at random is lower precision. Across most of the adjustments, the shapes of the distributions are similar and they are unbiased. A fairly uniform inflation of around 25% in the RMSE as a result of the lower response rates is evident.

Looking down the first column of charts and tables – MCAR, MAR 1 and NMAR – with similar overall response rates, the sizes of standard errors and shapes of the distributions do not vary so much, but the centres of the distributions move from the population mean for some methods – they become biased. For MAR non-response, methods such as the simple weighting class adjustments that do not take account of the sources of non-response are not successful at accounting for it, and become positively biased. Methods the incorporate those sources, i.e. factoring in prior performance and socio-economic background, are centred on the population mean. The measures of bias in these cases are close to zero. When data are NMAR, all methods tend to show some residual bias, but the capacity to remove any MAR components of the bias when using the latter adjustment approaches means that these methods keep the bias to a minimum.

With higher overall rates of non-response – the second column of charts and tables - the higher variance associated with the reduced overall sample size shows up in the larger standard errors and wider interquartile ranges. The effects on bias are similar as for the higher response rate scenarios. The combined effect of lower precision and bias contributes to much higher values for RMSE, up to 15-20, or 3 to 4 times higher than observed under full response.
The adjustment method that performed best and most consistently across all the non-response scenarios was the use of the generalised regression estimator, labelled as method ‘g’ in the charts and tables. This approach was the most precise of all the adjustment methods. The boxplots for this method had the smallest interquartile ranges and the standard errors under this method were substantially lower, even compared to standard estimators under full response. This method was also comparable with the other methods, such as the imputation methods and the estimation of response propensity, that took account of the auxiliary information in minimising bias. The approach performed well under all 1000 samples in each non-response scenario, with minimal instances of outlying values. As a measure of the success of this method in maintaining precision whilst minimising the extent of bias, the RMSE for this approach was at times around half the size of the other adjustment approaches. Even without considering the benefits of this approach with respect to managing non-response, the benefits of using such an estimator that makes good use of available auxiliary information related to outcomes are clear. With the additional benefit of protecting against possible non-response bias, the results from this approach present a clear case for inclusion in the estimation methods for Australian education surveys.

8.11.2 Shape, precision and bias: Estimating the population proportion - all scenarios

Figure 8-15 and Table 8-18 show the corresponding comparisons of distributions, precision and bias with respect to the estimation of the proportion of students achieving a defined benchmark of over 430 on the mathematics score.
The observations from these summary charts and tables are similar to those made with respect to the estimation of the population mean. Lower response rates are reflected with higher standard errors and wider interquartile ranges; adjustments that fail to account for the factors underlying non-response do less well, sometimes quite substantially so; methods that do successfully take those factors into account are more successful in reducing bias, although no method can completely reduce bias in scenarios where the underlying factors cannot be completely explained by available auxiliary information.

There appeared to be more volatility in outcomes when estimating the proportion compared to the mean. For example, the response propensity estimation approach (label e) produced a distribution of estimates that was centred slightly below the population proportion under MAR2, but under the NMAR scenarios, the distribution was centred above the population proportion.

The relatively poor performance of the weighting class adjustments that have not made full use of the available auxiliary data is revealed with the degree of bias displayed in the box plot summaries for these approaches, and also in the inflated values for RMSEs compared to the other approaches. Of the regression-based approaches, the multiple imputation approach (label i) was the most consistent at minimising bias whilst maintaining good precision. The use of generalised regression estimator (label g) also performed consistently well with good precision.
Figure 8-15: Summary distributions – estimating the population proportion – student level non-response
Table 8.18: Precision and bias – estimated proportion under student level non-response

<table>
<thead>
<tr>
<th></th>
<th>MCAR</th>
<th></th>
<th>MCAR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV</td>
<td>SE</td>
<td>bias</td>
<td>MSE</td>
</tr>
<tr>
<td>a</td>
<td>0.808</td>
<td>0.019</td>
<td>0.000</td>
</tr>
<tr>
<td>b</td>
<td>0.808</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>c</td>
<td>0.808</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>d</td>
<td>0.808</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>e</td>
<td>0.808</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>f</td>
<td>0.808</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>g</td>
<td>0.808</td>
<td>0.015</td>
<td>0.000</td>
</tr>
<tr>
<td>h</td>
<td>0.813</td>
<td>0.020</td>
<td>0.006</td>
</tr>
<tr>
<td>i</td>
<td>0.808</td>
<td>0.019</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MAR1</th>
<th></th>
<th>MAR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV</td>
<td>SE</td>
<td>bias</td>
<td>MSE</td>
</tr>
<tr>
<td>a</td>
<td>0.808</td>
<td>0.019</td>
<td>0.000</td>
</tr>
<tr>
<td>b</td>
<td>0.817</td>
<td>0.020</td>
<td>0.009</td>
</tr>
<tr>
<td>c</td>
<td>0.814</td>
<td>0.020</td>
<td>0.006</td>
</tr>
<tr>
<td>d</td>
<td>0.810</td>
<td>0.020</td>
<td>0.002</td>
</tr>
<tr>
<td>e</td>
<td>0.808</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>f</td>
<td>0.809</td>
<td>0.021</td>
<td>0.002</td>
</tr>
<tr>
<td>g</td>
<td>0.809</td>
<td>0.016</td>
<td>0.001</td>
</tr>
<tr>
<td>h</td>
<td>0.812</td>
<td>0.020</td>
<td>0.005</td>
</tr>
<tr>
<td>i</td>
<td>0.807</td>
<td>0.020</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NMAR</th>
<th></th>
<th>NMAR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV</td>
<td>SE</td>
<td>bias</td>
<td>MSE</td>
</tr>
<tr>
<td>a</td>
<td>0.808</td>
<td>0.019</td>
<td>0.000</td>
</tr>
<tr>
<td>b</td>
<td>0.846</td>
<td>0.019</td>
<td>0.038</td>
</tr>
<tr>
<td>c</td>
<td>0.834</td>
<td>0.020</td>
<td>0.026</td>
</tr>
<tr>
<td>d</td>
<td>0.824</td>
<td>0.021</td>
<td>0.017</td>
</tr>
<tr>
<td>e</td>
<td>0.820</td>
<td>0.021</td>
<td>0.012</td>
</tr>
<tr>
<td>f</td>
<td>0.823</td>
<td>0.021</td>
<td>0.015</td>
</tr>
<tr>
<td>g</td>
<td>0.825</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>h</td>
<td>0.827</td>
<td>0.020</td>
<td>0.019</td>
</tr>
<tr>
<td>i</td>
<td>0.818</td>
<td>0.020</td>
<td>0.011</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NMAR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV</td>
<td>SE</td>
</tr>
<tr>
<td>a</td>
<td>0.808</td>
</tr>
<tr>
<td>b</td>
<td>0.895</td>
</tr>
<tr>
<td>c</td>
<td>0.868</td>
</tr>
<tr>
<td>d</td>
<td>0.837</td>
</tr>
<tr>
<td>e</td>
<td>0.825</td>
</tr>
<tr>
<td>f</td>
<td>0.835</td>
</tr>
<tr>
<td>g</td>
<td>0.848</td>
</tr>
<tr>
<td>h</td>
<td>0.845</td>
</tr>
<tr>
<td>i</td>
<td>0.820</td>
</tr>
</tbody>
</table>
8.12 Summary of findings of student non-response investigations

The results from the investigation into strategies for managing student level non-response clearly show the differences between the different categories of missing data – MCAR, MAR and NMAR and the greater effectiveness of certain non-response adjustment approaches over other approaches in addressing non-response bias and precision.

When comparing outcomes of the student non-response investigations compared to the school non-response investigations of Chapter 6, one clear difference is that the distribution of estimates at the student level generally remained symmetric across the adjustment methods. At the school level, especially for higher rates of non-response and the more complex factors underlying that non-response, the distribution of estimates could be quite skewed. This was likely the result of the relatively small number of data points available (sometimes as few as 9) on which to base adjustments in the school non-response investigation. With more data points available under all scenarios at the student level, adjustment methods were less susceptible to this volatility.

When data are MCAR, there are few differences in results across the non-response management approaches. Even in the presence of quite high rates of non-response, most produce expected values for estimates of the mean and proportion at or near to the respective population values.

When data are MAR, simple non-response adjustment methods taking no account, or inadequate account of survey factors related to non-response, for example where
respondents from a school are simply weighted up to represent the number of students sampled from that school, could remain quite biased. Only methods that took account of all of the underlying factors associated with non-response could remove the overall bias. Across both versions of MAR investigated, some reduction in bias was observed with the formation of weight classes by sex compared to a simple adjustment weighting the respondents up to the sample size within school, but the outcomes remained biased unless the prior performance and socio-economic variables were factored into the adjustment. Amongst the other approaches to managing the non-response under the MAR scenarios, the ‘direct’ response propensity adjustment to the design weights appeared to be consistently slightly biased compared to the regression based and imputation approaches. There appeared to be more instability in outcomes for the second MAR scenario with higher non-response. Under MAR2, the method involving post-stratification by classes defined by response propensity produced slightly upwardly biased outcomes and the regression estimator approach with weights adjusted by sex within school tended to slightly underestimated the mathematics mean (Table 8-11). The weight diagnostics indicated a small increase in weight variation compared to the MCAR scenario, with weighting effects up to around 1.1 although the variation in weights was still reasonably minimal.

As expected, under each of the NMAR scenarios, some bias remained under every non-response management approach used. As with the MAR methods, those that did not take account of factors related to response performed less well than those methods that took the factors related to response into account. With the relatively high overall response rate of 77% explored in the first NMAR scenario, most of the regression and imputation based methods performed quite well in reducing the bias, without substantial
inflation of standard errors. In the second NMAR scenario, where the non-response related to mathematics outcome was more uniformly distributed across the performance and socio economic background factors related to non-response, the methods performed better at reducing the bias, even though the response rate had dropped to around 55%. There was more variability in outcomes across the methods in this scenario. Once again, the single regression imputation performed less well under the lower response rate. Of all the methods used under this scenario, multiple regression imputation showed the least bias with respect to estimating the mathematics mean, but at the expense of a slightly larger standard error. When estimating the benchmark proportion, the multiple regression imputation was also less biased, and the standard error was slightly lower than for other methods.

With the stronger rates of non-response in the third NMAR scenario, and with non-response more varied across performance and socio-economic categories, there was more variation in outcomes across the methods. Under this scenario, the best performing method in terms of reducing bias was the use of the regression estimator, with design weights adjusted for non-response by sex within school. Once again the multiple imputation approach was also quite successful in reducing bias.

In the NMAR scenarios, where not all drivers of non-response could be used in adjustments, bias remained, but it was limited to those factors. Any MAR component related to non-response in these scenarios could still be addressed using several of the methods.
It was clear overall that the methods that were able to take account of all of the main factors underlying non-response performed well in reducing non-response bias in survey outcomes. With clear non-response patterns observed in Australia’s participation in TIMSS in Chapter 5 against key variables including socio-economic background and prior academic performance, the outcomes from this chapter can be used to justify the efforts to obtain such variables for use in non-response adjustments from available data sources such as NAPLAN.

The comparison of RMSEs across the adjustment methods in Table 8-17 and Table 8-18 gave a strong indication of the relative quality of the method. For the estimation of means the regression estimator has appreciably lower RMSE than any estimator except in the case of NMAR with very low response rates. This is mainly due to the improvements in precision indicated by the lower standard error. Only in the case of the stronger NMAR scenarios with higher bias does the RMSE for the regression estimator move slightly higher than the imputation methods. When estimating the proportion the lower standard error of the regression estimator tends to also apply, but the gain over other methods is not as strong. It still performs well in the MCAR and MAR scenario but for the NMAR scenarios the lower bias of the multiple imputation approach leads to lower RMSEs.

More generally across the different types of non-response management strategies, it was evident that the regression based approaches – e.g. the use of a regression estimator, or single or multiple imputation – consistently performed well compared to other methods. The method involving the generation of estimated response propensity appeared slightly less robust, with outcomes from this approach sometimes appearing more biased.
Chapter 9  Summary and conclusions

9.1 Non-response and the quality of survey outcomes

Educational surveys are an important mechanism by which policy makers monitor the progress of Australian school systems in delivering high quality education for students. There is a stated policy of using educational surveys to improve outcomes, and this is reflected in the considerable amount of survey activity that is undertaken by Australian Commonwealth State and Territory governments, particularly through the National Assessment Program (NAP).

As with many other countries, and in many other social science domains, Australian educational surveys experience a level of non-response. Non-response can affect the quality of survey outcomes, with potential losses in precision and bias effects. As noted in section 1.1 the loss of precision occurs directly and indirectly through the failure to
collect data from sampled units. The risk of non-response bias arises when respondents and non-respondents differ with respect to the outcome measures of the survey.

School participation in surveys conducted within the NAP is effectively mandatory, and therefore there is minimal non-response of schools in these surveys. For surveys of Australian school students that are not within the NAP, school non-response does occur, an example was provided in section 6.1. Non-response is also experienced at the student level, both for NAP and for other surveys.

The challenge of running a successful education system in a country as diverse as Australia is an enormously complex and costly endeavour, but because of the perceived rewards in terms of economic and social benefits that come with a well-educated populace, one that Australian governments have shown a strong willingness to invest in. Survey work which has been put in place to monitor the progress in improving these systems should involve the highest quality statistical methods in all aspects of their implementation. This research considers whether data that is already being collected in the Australian context could be better utilised towards one important aspect of survey work, the management of that non-response.

9.2 Minimising non-response during the data collection phase

An important component of quality survey practice is to address the issue of non-response and its potential effects on survey outcomes. This includes developing survey operations to minimise the degree of non-response as well as applying statistical methods to adjust for the incidence of non-response in the collected data. Operational
factors will include activities like careful work with communicating with schools about the importance of the survey, minimising the burden on schools arising from the survey, for example through the provision of external personnel to administer the survey, and following up with schools and students who have not responded. These were discussed in section 2.4.

9.3 The approaches used to manage non-response

As canvassed in the literature review, a range of statistical methods are used to address any non-response that has occurred in a survey to minimise potential biases and loss of precision in key estimates arising from the non-response. These methods include weighting class adjustments, post-stratification, the use of estimators that take into account auxiliary data related to outcomes, response propensity methods, and regression and imputation methods. The availability of data external to the survey that is related to outcomes and / or patterns of response can be a substantial help in addressing those non-response effects. Of particular interest is whether the data collected through the annual NAPLAN census of student literacy and numeracy, data which has to this point not been made available for the management of non-response for other NAP surveys, could be used to improve the methods used to address non-response in sample-based surveys.

9.4 Understanding the factors relating to achievement

The first step in an investigation into how best to address the effects of non-response in a survey is to examine the factors that are important in explaining achievement outcomes and also response patterns. To this end, in chapter 4, an investigation of the TIMSS 2011 survey of mathematics and science achievement was conducted, and found
a number of important factors explaining variation in achievement on this survey. Of the factors investigated, an alternative measure of performance contributed most strongly, approximately doubling the proportion of variance explained in achievement outcomes from the survey (section 4.4.3). Also important were measures related to the socio-economic background of the student. Other factors that were related to outcomes included school climate, school type, location, and student’s enjoyment and confidence in the subject area, although conditional on prior performance measure and socio-economic background, these additional variables contributed a relatively small degree of additional explanatory power. The outcomes of these investigations were consistent with findings of previous Australian research summarised in section 3.3 into the factors related to achievement.

9.5 Understanding the factors driving non-response

Chapter five extended the investigation of the TIMSS 2011 dataset, to examine response rates for different sectors of the population to examine factors relating to response. It found clear evidence that response rates declined as measures of student performance including the NAPLAN school mean, and also measures of socio-economic background declined. As with the factors explaining performance, response rates varied for groups defined across a number of variables including the school location, the proportion of students from Aboriginal and Torres Strait Islander backgrounds, and the school attendance rate.
9.6 Non-response in a two-stage sample design

With the evidence that prior performance, and socio economic background related variables are important in explaining both achievement outcomes and response, the remainder of the thesis explored how such data might be incorporated into non-response management.

The investigation was broken up into two major components, reflecting the two stages of selection – schools, then students from those sampled schools – that is intrinsic to the sample design for all major Australian educational surveys. For each stage, the focus was on the effects at that stage alone. 100% response rates and perfectly unbiased estimates were assumed with respect to the other stage. Initial assumptions driving the model of response were as far as possible based on observations from the analysis of the TIMSS data in Chapter 4. For both the school and student non-response investigations, the starting point were the response rates across key performance and socio-economic measures that were observed from TIMSS.

9.6.1 School level non-response

For the investigation of school non-response it was important to investigate management methods under realistic scenarios for surveys experiencing school level non-response. However there was very little school non-response observed for the TIMSS survey as school participation in this, as with all NAP surveys is essentially mandatory. In the absence of evidence of the drivers of school non-response from the TIMSS survey the starting point for these investigations was to assume that a school’s propensity to respond would be proportional to the response propensity of the population of students the school caters to, as observed from TIMSS. As noted (section

315
6.7.2), there may well be other drivers of non-response at the school level such as the burden of participation, and further research into this area would be of great value, but this was considered a reasonable starting point for this part of the investigations within this thesis.

From these starting assumptions, investigations were conducted into the relative effectiveness of non-response management approaches under a range of different scenarios, for example with higher rates of non-response, and with non-response more strongly related to outcomes. As noted in section 6.1, surveys of Australian school students that are conducted outside the NAP encounter much lower rates of response at both school and student levels. Not all educational surveys within Australia have the benefit of the resources and full policy backing of the NAP surveys. It was important to explore the robustness of methods under a range of scenarios involving higher rates of non-response and differing patterns of missing data.

The school non-response investigation was able to use of the full population of Victorian government schools by making use of the Sampling Frame prepared and updated annually by the Australian Council for Educational Research (ACER) using data from the Commonwealth State and Territory governments, as well as data published on the MySchool website including average school performance, and socio-economic background measures, in particular the Index of Community Socio-educational Advantage (ICSEA).

When non-response was high the distribution of non-response adjusted outcomes could be very skewed. When the non-response was MCAR, the mean of the
distributions was generally close to the population mean, but with quite a number of estimates substantially higher than the population figure, pushing the mean some distance from the median value. A good example of this was Figure 6-10 showing the distributions of estimates of the proportion of students in the top numeracy band under high levels of MCAR non-response. The regression and imputation based methods had less extreme observations and had more symmetric distributions. The variation in outcomes was lower, as indicated by the lower standard errors.

Under scenarios with data that was missing at random – for example in estimating the mathematics mean under ‘mild’ MAR non-response (Figure 6-11), the skewness of distributions – particularly for weighting class and post-stratification adjustments – was still very evident when non-response was high. The standard errors for those methods were once again higher than for the regression and imputation based methods. The methods involving estimation of response propensities performed less well, suggesting instability in model estimation with as few as 9 records used in the model estimation. The regression estimator and the school substitution methods performed well relative to these other methods. With a stronger degree of missingness associated with factors such as prior performance, management methods that did not take those factors into account could quite substantially overestimate the population parameter, for example the non-response adjustment by school location (method e) in Figure 6-12.

When a component of the school non-response was related to outcomes in a way that could not be fully explained by the explanatory variables, i.e. NMAR, bias was generally present in all of the management approaches. However, it was clear that the regression estimator and school substitution were more successful in minimising bias,
and also maintaining good precision in the estimation (Figure 6-13, Figure 6-14). The weighting class and post-stratification approaches under these scenarios could produce quite extreme outcomes in some cases.

Overall it was clear from the investigation into school level non-response that the primary method used to manage non-response in NAP surveys – the use of school substitution – performed well compared to other methods of managing non-response, at least with this subpopulation of Australia’s education system. In line with the study of the use of substitution in school based surveys discussed in section 2.6.5, the use of school substitution did not completely remove non-response bias, but performed at least as well as any of the other methods at managing non-response, and in fact better than many of those other methods. The standard error under school substitution was similar to that achieved with full response because the number of data points used in the estimation was the same.

The school substitution approach used in NAP surveys (including in the major international surveys such as TIMSS and PISA discussed in this thesis) limit the identification of potential substitutes to two, usually the two adjacent schools on the sorted list (section 6.9.7). In practice it is sometimes the case that neither the sampled school nor one of its replacements participates in the assessment. In this case, amongst the other management methods available for school level non-response, the use of a regression-based estimator appears most likely to produce an outcome with least bias and most precision. Weighting class estimates and post-stratification were least successful in managing non-response, sometimes producing very biased outcomes. Response propensity estimation appeared somewhat unstable given the relatively small
sample size in the context of school participation. The precision of estimates under simple regression-based imputation was likely underestimated in the approach used in the investigation. The development of more sophisticated regression-based and yet stable imputation models would be unlikely, especially with higher levels of school non-response. In summary, at least for this part of the Australian education system, the use of school substitution supplemented as needed with the use of a regression-based estimator would seem the approach most likely to obtain an estimate consistent with the population characteristic.

Victoria has a large population and has less geographic diversity than other parts of the country. For other parts of the Australian population of students, school substitution may prove less successful in terms of finding good substitutes for the sampled, non-responding schools and students. The methods outlined in this thesis could be modelled across all jurisdictions to find methods that appear to produce the most stable estimates. Based on the outcomes from the investigations in this thesis, they are likely to be drawn from a combination of school substitution and regression based estimation.

9.6.2 A database for exploring student level non-response

In the preparation for the investigation into student level non-response, an innovative part of the investigation undertaken was the use of donor-imputation (section 2.6.5) to produce a simulated population of students as the basis for analysis. In fact Australia does have, through its NAPLAN assessment, a comprehensive database of student and school performance with school level data published now over many years via the MySchool website. However, the student level data were not made available for the purposes of this research, and alternative methods were needed to produce a dataset that realistically represented the diversity of the student population. By making use of actual
student records from schools identified as near neighbours with respect to prior performance and socio-economic background the complexity of factors that contribute to student outcomes was retained. A similar approach could be adopted for any country participating in a large scale survey such as TIMSS, where such data was either not available, or does not exist within the national context. The further step of factoring in a layer of clustering of students within classes within schools also contributed to the quality of the simulated dataset. Future investigations might look to explore more sophisticated approaches towards identifying the most suitable near neighbours within which to identify donors to produce a realistic simulated population.

9.6.3 Student level non-response

As with the investigation of school level non-response, the student level non-response investigation began with models of non-response that corresponded with the observed response rate patterns from the TIMSS survey. Unlike at the school level, where factors contributing to non-response were only conjectured, at the student level, it was possible to use rates observed in the TIMSS survey as a starting point for investigations. Further scenarios involving higher rates of non-response, and non-response more or less closely related to outcomes stemmed from these original investigations. Non-response adjustment methods such as unit substitution that were used at the school level were not explored at the student level for the reasons discussed in section 2.6.5.

The distributions of outcomes in the student non-response models explored tended to be more symmetric than those observed under school level non-response. Even with high rates of non-response, it appeared that there were sufficient numbers of responding records to produce more stable estimates. For the MCAR models, the most notable feature across the models was the relatively higher standard errors observed in the
weighting class adjustment approaches compared to other approaches. For example in
the case of 35% overall response rates with MCAR induced non-response – summarised
in section 8.10.3 (estimating the mean) and 8.10.4 (estimating the proportion above the
benchmark), the standard errors were around 50% higher than those from the full
sample estimate of the population mean for the weighting class adjustments, but around
25% higher for the regression and imputation based methods (Table 8-8). A key
component of the increase in standard error was the drop in sample size, with standard
type errors inversely proportional to the square root of the sample size. Beyond that factor,
differences in standard errors could be compared across adjustments made from the
same sized samples and the regression and imputation methods performed better in this
respect.

When data were MAR, it was notable that the adjustments that did not fully take into
account the factors behind the non-response could produce biased estimates, even with
relatively high rates of response. For example the weighting class adjustments used in
estimating the population proportion explored under MAR response in section 8.10.4
with distributions shown in Figure 8-9. The estimates based on simple weighting class
adjustments that did not take account of the factors underlying non-response (labels b
and c) remained biased. The effect was even more noticeable when there was a stronger
relationship between missingness and survey variables which was not taken into
account in the adjustment methods, for example, adjustments ‘b’ and ‘c’ under the
MAR2 scenario as displayed in Figure 8-10.

As with the school non-response analyses, under a scenario where missing data was
related to outcomes in ways that could not be explained by other survey variables, no
method was able to successfully remove all of the bias, but some methods were more successful than others. In general the regression based approaches performed better both with respect to minimising residual bias and also maintaining low standard errors. In some cases under these scenarios the simpler approaches such as weighting class adjustments and post-stratification had little or no success in reducing bias, and were much worse with respect to precision.

When data were not missing at random, while no method was able to remove the underlying bias, the weighting class and post-stratification adjustments generally performed quite poorly compared to the regression and imputation based methods, for example Figure 8-12 and Figure 8-13.

With the demonstrated patterns of non-response across a measure of prior performance observed in the analysis of the TIMSS data in chapter 5, for example the charts of participation rates against categories of the school NAPLAN numeracy mean and the TIMSS science performance measure in section 5.4, it is reasonable to conclude that non-response on a survey of academic performance such as TIMSS will be to some degree related to the outcomes, and that there will be some associated non-response bias. Even when there is a relatively good overall response rate, that bias can lead to estimated outcomes some distance from the population mean, especially when the bias has not been addressed by factoring in the most likely variables associated with response and with outcomes. For example in the first (and mildest) NMAR model investigated in the student non-response investigation, student response rates were 77%, but for the most simple weighting class adjustments the expected value of the population mean was more than 10 score points, more than two standard errors from the
true mean (Table 8-12). The mildest NMAR model investigated at the school level with an 80% school response rate, the estimated mean with no non-response adjustment – effectively the approach currently used with respect to school level non-response for a particular stratum of TIMSS – was five score points higher than under full school response (Table 6-7). To put that into some perspective, a drop of 10 scale score points on the Year 8 mathematics outcome for Australia overall in the TIMSS 2011 survey would correspond to a drop in the country ranking from 12th to 15th. (Of course similar or greater bias effects might also be present in other participating countries that achieved a higher or lower ranking.)

9.7 Conclusions and further investigations

It was clear from the investigations that making use of variables known to be correlated with outcomes, particularly prior performance and socio-economic background helped to reduce the effects of non-response bias, and to maintain good precision. They present a strong case that, at least with respect to this part of the Australian school system, management of non-response for Australian educational surveys would be enhanced with the use of the data available through NAPLAN and MySchool.

The use of school substitution was confirmed in the investigations in this thesis as a worthwhile measure against potential non-response bias at the school level. Given the degree to which auxiliary data on prior performance and socio-economic background were successful in reducing the effects of non-response bias - observed in both the school and student level investigations - the approach of school substitution would be enhanced if these factors were more directly incorporated into sample design and stratification for Australian student surveys, for example through the use of categories.
of prior performance and socio-economic variables as stratification variables in the sample design. By so doing, schools on the sampling frame adjacent to sampled schools, i.e. the schools assigned as potential substitutes, would have similar prior performance and socio-economic backgrounds as the sampled school, and the benefits demonstrated in the investigations in this thesis of making use of this auxiliary data in terms of managing potential non-response would be more directly incorporated into the school substitution approach.

It is not uncommon in Australian educational surveys to have instances where neither the sampled school nor the schools assigned as potential substitutes participate in the survey, so school substitution does not represent a complete solution to the management of school-level non-response. The regression and imputation based adjustments that made use of auxiliary information on prior performance and socio-economic background at the school level investigated in this thesis also performed demonstrably better than the simple stratum level adjustments that are currently in use for the management of school non-response.

At the student level the investigations in this thesis clearly show that the simple adjustments that are currently being used to address survey non-response in Australian educational surveys, based on survey design variables such as State, sector and location, are ineffective in addressing non-response arising from the variables that most explain outcomes in Australian educational surveys, prior performance and socio-economic background. Even under fairly mild conditions of non-response, such as the NMAR1 conditions shown in section 8.10.6, the estimated outcomes derived from simple non-response adjustments could be very biased. The outcomes from the regression and
imputation based methods that take account of prior performance and socio-economic background, while still slightly biased, performed much better under these conditions. Under stronger conditions of non-response, such as those explored in section 8.10.9, estimates derived following the simpler non-response adjustments could remain seriously biased.

Another clear observation of the investigations was that the precision of estimates was reduced when auxiliary data related to the key outcomes of the survey was not incorporated into the estimation. In contrast, the use of the generalised regression estimator incorporating prior performance and socio-economic background produced notably more precise estimates of outcomes, even than when under full response using the standard estimators. On the basis of the precision of estimates alone, estimation of outcomes would be enhanced with methods that incorporate these auxiliary variables. The fact that these approaches were also protective of potential non-response bias, even under quite strong non-response conditions, speaks very strongly for the incorporation of these methods into the estimation of outcomes.

9.7.1 Recommendations for future Australian educational surveys

On the basis of the investigations of this thesis, I make the following recommendations for future Australian educational surveys:

1) The census-based NAPLAN prior performance data (in particular) and also student-and school level socio-economic background data, such as the school level ICSEA measure and the student level components that make up that measure (section 3.2.2) should be made available to assist with sample design, non-response management and estimation for Australian educational surveys.
a. The incorporation of these data into stratification at the sample design stage will improve the method of school substitution that the thesis has shown to be a successful strategy for managing school-level non-response.

b. These auxiliary variables have shown through the investigations in this thesis to produce more precise estimates. This was particularly demonstrated with the use of the generalised regression estimator in the investigations undertaken at the student level.

c. These auxiliary variables were also shown to be good protections against the effects of non-response bias, at both the school and student levels. Methods which did not make use of these auxiliary variables could remain seriously biased.

2) Investigations should be conducted along the lines of those explored in this thesis, making use of the NAPLAN census-based data to confirm the results of the investigations based on simulated data in this thesis. Investigations should be extended to other States and sectors of the Australian population to examine whether results that were observed for the sub-population explored in this thesis also apply to other States, sectors and locations.

3) Subject to the outcomes of the investigations using NAPLAN data and extended to other States and sectors, it is likely that management of non-response at the school level should involve a combination of school substitution and the incorporation of auxiliary data into the estimation of outcomes.

4) The management of non-response at the student level should make use of auxiliary variables available through the NAPLAN and MySchool census data collections. The investigations in this thesis point to generalised regression
estimation as the best performing approach, with the most precise estimates and good protection against student level non-response bias. As was observed in Micklewright et al. (2012), (section 2.8) the capacity to make use of auxiliary data at the population level provides an advantage over other approaches in that it can capture non-response effects happening at both the school and student levels. The relative merits of this approach compared to the other methods explored in this thesis can be examined across a range of States and sectors making use of the NAPLAN data.

9.7.2 Technical issues for further investigation

In the research undertaken for this thesis, aside from the discussion about the weighting applied in the management of non-response in the TIMSS survey (section 4.3), there was no attempt to conduct a joint 2-level investigation into the management of non-response in educational surveys. Each stage of selection was considered separately. This was partly because there was no strong basis for modelling school level non-response in the Australian context, because the surveys conducted within the NAP effectively mandate school participation response rates are generally between 95% and 100%. Also, while the research sought to draw from observations of response rates appearing in the student population as its starting point for school level models, the observation was made (section 6.7.2) that the factors driving school level response may be quite different to those that exist at the student level.

As noted in section 5.4 a useful analysis for further research would be to examine the interactions between variables and their effect on participation rates through regression modelling, to obtain better estimates of the relative effects of these variables.
As observed in section 2.7.4, the availability of auxiliary data at the population level in
the use of a generalised regression based approach to non-response management has the
capacity to address non-response effects across the population, at both school and
student levels. However further modelling of the drivers of non-response at the school
level may enable supplementary approaches, for example through the identification of
further factors driving non-response at this level, and auxiliary data related to that.

In other contexts, for example other countries, a stronger basis for modelling school
level non-response and for the joint non-response patterns across the two levels might
be available. In this case additional methods such as those involving regression
estimation for a two-stage design (see for example Särndal et al. (1992)) would be a
valuable extension to the research in this paper. Regression estimation can be extended
to use a mixed model as described in Park and Fuller (2009)
References


Appendix A: Diagnostics of multiple regression explaining TIMSS mathematics achievement
A.1 Regression Model 1: Mathematics achievement explained by Sampling Frame variables

*Year 8 Mathematics achievement explained by State, sector, location, school type, school size, SEIFA IEO measure for the school, proportion of girls, proportion of ATSI students*

The fit diagnostics for the model making use of sampling frame variables (Fig A 1) indicate the assumptions behind linear regression have broadly been met. The plot of studentised residuals versus predicted values shows the majority of data points within...
+/- 2, with relatively few points outside these boundaries. At the top end of this plot there appears to be a small number of anomalous schools, particularly one where the predicted value is high but the residuals are strongly negative. This school was more clearly identified in the plot of residuals against Indigenous student percentage (figure A2). The school is made up entirely of students with Indigenous backgrounds who overall have performed less well than their characteristics across the other variables in the model would suggest. Were further predictions to be based on this model, consideration might be given to dropping this school from further analysis.

A.2: Residuals on key variables: regression model 1 explaining mathematics achievement

The normal quantile-quantile (QQ plot) and the normal density plot indicate few departures from normality in the model.

A.2 Regression Model 4: Mathematics achievement explained by Sampling Frame variables as well as MySchool and TIMSS variables

Year 8 Mathematics achievement explained by sampling frame variables (Model 1) plus proportion of LBOTE students, School Numeracy mean, attendance rate) and TIMSS variables (Home Educational Resources, student sex, Science performance)
The fit diagnostics under Model 4 show some improvement compared to Model 1 with respect to fewer large residuals and a better fit of the observed values compared to predicted values. There do not appear to be data points with excessive influence in the model. The Q-Q plot and distribution of residuals appear normal.
A.3 Regression Model 5: Model 4 with additional TIMSS variables related to school climate (School Emphasis on Success / School Discipline and Safety) and student affect (Like Maths / Confidence in Maths / Engaged with Maths).

As with Model 4, Model 5 shows no major concerns with respect to the assumptions underlying the multiple regression model.