Research Report DCSF-RW086

A Strategy for Handling Missing Data in the Longitudinal Study of Young People in England (LSYPE)

Andrea Piesse and Graham Kalton
Westat

department for
children, schools and families

Research Report No DCSF-RW086

A Strategy for Handling Missing Data in the Longitudinal Study of Young People in England (LSYPE)

Andrea Piesse and Graham Kalton

Westat

ISBN 978 1 84775 337 3

Table of Contents

Chapter			Page
Executive Summary			ii
1		Introduction	1
2		Why Missing Data Adjustments Are Necessary	3
3		Imputation vs. Weighting	6
	3.1	General Considerations	6
	3.2	Choosing Between Weighting and Mass Imputation for Partial Non-response	6
4		Weighting for Missing Waves and Components	9
	4.1	Component Combinations of Significant Analytic Interest	9
	4.2	What Constitutes Component Response?.	10
	4.3	Missing Data in Wave 1	11
	4.4	Missing Data in Waves 1 and 2	12
	4.5	Missing Data in Waves 1, 2, and 3	13
	4.6	Recommendations for Weighting	16
5		Imputation for Item Non-response	18
6		Guiding Users on Statistical Analyses	23
References			24

Executive Summary

The Longitudinal Study of Young People in England (LSYPE) is a panel survey of young people and, initially, their parents. Beginning in 2004, the sampled young people (from the cohort born between 1 September 1989 and 31 August 1990) are being surveyed annually for ten years or more. For the first four waves, interviews were also attempted with at least one parent or guardian living with the sampled young person. For these waves, in addition to the data collected in face-to-face interviews, the sampled young people and the main parents were asked to enter their responses to a number of more sensitive items directly into the interviewers' computers using a computer-assisted self-interview mode of data collection. Beyond Wave 4, the survey uses a mixed mode approach with web-based interviews and telephone interviewing, as well as some face-to-face interviews. The data collected in the survey are also merged with administrative data from the National Pupil Database and the Pupil Level Annual School Census.

As a result of the longitudinal survey design, the multiple respondents, and the different data sources, there are many possible types of non-response. As in any longitudinal survey, there is non-response at the first wave and further non-response at the second and subsequent waves. In addition, partial non-response occurs at each wave when data are not collected for one or more components (young person interview, one or other parent interview). There are also the item non-responses that occur when acceptable responses are not obtained for one or more of the items in a component that is otherwise complete.

Some form of compensation is needed to reduce the biases in survey estimates resulting from missing data. Weighting adjustments and imputation are alternative forms of general-purpose compensation procedures for handling missing survey data. Both these procedures employ an assumption that the missing data are missing at random within subsets of the sample. Carefully developed compensation procedures based on this assumption should generally reduce the biases in survey estimates that occur when no adjustments are made for the missing data. It needs to be recognised that there are significant costs involved in implementing these procedures. However, once the procedures are applied to the survey data set, any analyst can analyse the data in a routine way. The application of these procedures makes it unnecessary for each analyst to develop his or her own compensation procedures or to analyse the respondent data without compensation.

The aim of this report is to propose methods to compensate for various sources of missing data in the LSYPE. Data from Waves 1 to 3 of the survey have been analysed to inform the development of a missing data strategy. Patterns of non-response have been examined to determine the exact nature of the missing data with respect to sample attrition (wave non-response), missing components (among wave respondents), and item non-response.

The two standard methods for handling missing survey data are weighting adjustments and imputation. The advantage of weighting is that it maintains the associations between all the items in the respondent data set. When applied with partial non-response, however, weighting inevitably loses some of the partial information that is available. With imputation, missing responses are assigned values, with all the recorded responses retained, thus making full use of the reported data. The dominant concern with imputation is that it may affect the associations between items. Based on the consideration of several factors, including the retention of reported data to the extent possible, the degree to which associations among variables can be preserved,

the degree to which bias can be reduced or avoided, and transparency and simplicity of use, we recommend the use of imputation to handle item non-response and weighting adjustments to handle unit and component non-response.

In order to retain as much of the reported data as possible for combinations of components that are of major analytic interest, while keeping the number of sets of weights computed to a manageable level, 15 different sets of weights are proposed for the first three waves of the LSYPE (see Table ES-1). Eight sets of weights are intended for cross-sectional analyses of data from individual waves, and seven are selected for longitudinal analyses of data derived from more than one wave. Analysts should identify the sets of weights that contain all the components needed for their analyses and choose the set that has the largest sample size.

With respect to item non-response, ideally all the missing values would be imputed. However, careful imputing for all missing data items within completed components of the LSYPE would be an enormous task because of the size and complexity of the survey. The proposed approach is therefore to impute only for selected items that are likely to be key for many analyses.

The following important topic areas in the LSYPE data set have been identified as candidates for imputation: household income, highest parental qualifications, family's National Statistics Socio-economic Classification (NS-SEC), attitudes and aspirations towards schooling, and bullying. In the future, other items may be identified for imputation based on considerations of cost and benefit. To take advantage of and maintain the strong associations between items related to income, parental qualifications, and NS-SEC, we recommend that they be imputed collectively.

Table ES-1 Recommended sets of weights for Waves 1 through 3 of the LSYPE

Weight description	History data required *	Wave 1 data required	Wave 2 data required	Wave 3 data required	Sample size
CROSS-SECTIONAL					
Wave 1		Young person only			15,298
Wave 1		Young person and main parent only			14,763
Wave 1		Young person, main parent, and second parent			13,552
Wave 2			Young person only		13,239
Wave 2			Young person and main parent only		12,852
Wave 2			Young person, main parent, and second parent		11,722
Wave 3				Young person only	12,243
Wave 3				Young person and main parent	11,893
LONGITUDINAL					
Waves 1-2		Young person only	Young person only		12,993
Waves 1-2	Yes	Young person and main parent only	Young person only		12,267
Waves 1-2		Young person only	Young person and main parent only		12,621
Waves 1-2	Yes	Young person, main parent, and second parent	Young person, main parent, and second parent		10,582
Waves 1-3		Young person only	Young person only	Young person only	11,866
Waves 1-3	Yes	Young person and main parent only	Young person only	Young person only	11,241
Waves 1-3	Yes	Young person, main parent, and second parent	Young person only	Young person and main parent only	10,257

* Relationship data provided at either Wave 1 or Wave 2 by a resident natural parent.

We also recommend that the imputation strategy focus on filling in values for individual survey items to provide analysts with maximum flexibility in computing derived variables for their specific analyses. Measures that are reasonably stable over time, such as parental qualifications, and those for which the best information has likely already been collected in the first three waves of the survey, may be imputed once. However, for other items, there may be a need for initial imputation as the data for a given wave are being processed, followed by re-imputation as later survey data become available.

1 Introduction

The Longitudinal Study of Young People in England (LSYPE) is a panel survey of young people and, initially, their parents. Beginning in 2004, the sampled young people (from the cohort born between 1 September 1989 and 31 August 1990) are being surveyed annually for ten years or more. For the first four waves, interviews were also attempted with at least one parent or guardian living with the sampled young person. For these waves, in addition to the data collected in face-to-face interviews, the sampled young people and the main parents were asked to enter their responses to a number of more sensitive items directly into the interviewers' computers using a computer-assisted self-interview (CASI) mode of data collection. Beyond Wave 4, the survey uses a mixed mode approach with web-based interviews and telephone interviewing, as well as some face-to-face interviews. The data collected in the survey are also merged with administrative data from the National Pupil Database and the Pupil Level Annual School Census (PLASC).

As a result of the longitudinal survey design, the multiple respondents, and the different data sources, there are many possible types of non-response. As in any longitudinal survey, there is non-response at the first wave and further non-response at the second and subsequent waves. In addition, partial non-response occurs at each wave when data are not collected for one or more components (young person interview, one or other parent interview). There are also the item non-responses that occur when acceptable responses are not obtained for one or more of the items in a component that is otherwise complete.

The aim of this report is to propose methods to compensate for various sources of missing data in the LSYPE. The recommendations are based on the first three waves of data. However, the intent is to select methods that can be applied over time as more waves of data become available. The procedures should enable reproducible analyses of the data, mitigate the potential bias from missing data, and produce data sets that are simple enough for analysts of various levels of sophistication to use. It needs to be recognised that there are significant costs involved in implementing these procedures. However, once the procedures are applied, any analyst can analyse the data in a routine way. The application of these procedures makes it unnecessary for each analyst to develop his or her own compensation procedures or to analyse the respondent data without compensation.

The two standard methods for handling missing survey data are weighting adjustments and imputation. In simple cross-sectional surveys, the choice between them is typically straightforward: Weighting adjustments are used to compensate for unit non-response, and imputation is used to compensate for item non-response. In complex surveys like the LSYPE that experience various types of wave and component non-response, the choice between weighting and imputation for handling the partial non-response is less clear. The choice hinges on several factors, including the retention of reported data to the extent possible, the degree to which associations among variables can be preserved, the degree to which bias can be reduced or avoided, and transparency and simplicity of use.

Data from Waves 1 to 3 of the LSYPE have been analysed to inform the development of a missing data strategy. Patterns of non-response have been examined to determine the exact nature of the missing data with respect to sample attrition (wave non-response), missing components (among wave respondents), and item non-response. Summary tables are provided in this report to support the recommendations.

A critical consideration in developing a strategy for handling missing data in the LSYPE is to take account of the needs of analysts. For example, which items and which subsets of waves and components are of greatest importance? A number of researchers who are familiar with the LSYPE through their work in analyzing the study's data have been consulted to obtain information about user needs. From these discussions some common themes were identified. Westat thanks the following people for making time available to discuss their experiences with missing data in the LSYPE, to identify important survey components and key items, and to provide their views on future research directions using this data source:

- Claire Baker, Department for Children, Schools and Families
- Claire Crawford, Institute for Fiscal Studies, London
- Michael Greer, Department for Children, Schools and Families
- James Halse, Department for Children, Schools and Families
- Ian Noble, Department for Children, Schools and Families
- Steve Strand, Institute of Education, University of Warwick

Before consideration of the alternative methods to compensate for missing data in the LSYPE, we first discuss reasons why simply analyzing the complete data is problematic.

2 Why Missing Data Adjustments Are Necessary

A notable strength of the LSYPE is that the study is based on a nationally representative probability sample of young people. As a result, the study's data can be analysed to produce estimates for all young people in the cohort born between 1 September 1989 and 31 August 1990 and for any subset of this cohort (e.g., subsets defined by ethnic group, geographic region, or socio-economic classification). Even if there were no missing data, survey weights would be needed to produce valid estimates for all young people in the cohort or for all those in a given subset. A young person selected with a probability of 1/*k* represents *k* young people in the cohort and hence is counted *k* times in the analyses. This process is achieved by assigning the young person a base weight of *k*. In this way, weighted totals computed from the sample provide estimates of the corresponding totals for the full cohort. If all the young people were sampled with equal probabilities, the base weights would not need to be used in estimating average quantities such as percentages, means, and correlation and regression coefficients because these quantities are all ratios of totals in which the base weights would cancel out. However, in the LSYPE, maintained schools with a high proportion of pupils in receipt of free school meals were over-sampled by a factor of 1.5, and young people from major minority ethnic groups were over-sampled at the pupil level. As a result, the base weights should be used in any analyses, even for estimating average quantities for all young people in the cohort.

With complete response, the use of base weights guarantees approximately unbiased estimates of cohort parameters. If the survey data were analysed without using the base weights, the resultant estimates would likely be biased. If those young people with higher weights - i.e. those with lower selection probabilities - have different characteristics from the others, estimates of average quantities will be biased.

The above discussion deals with the situation where no data are missing for sampled young people. Analyses that use only the base weights will lead to biased estimates when data are not available for some of the sampled young people. The only case in which the estimates of average quantities will be unbiased occurs in the improbable situation when the missing data are missing completely at random (MCAR) with respect to the variables involved in the analysis (Little and Rubin, 2002). In other cases, some form of compensation for the missing data is needed to reduce the biases in the survey estimates. Weighting adjustments and imputation are alternative forms of general-purpose compensation procedures for handling missing survey data. Both replace the unrealistic MCAR assumption by a missing at random (MAR) assumption, that is, that the missing data are MCAR within subsets of the sample. While the MAR assumption is also generally false, it is a more realistic assumption. Carefully developed compensation procedures based on the MAR assumption should generally reduce the biases in survey estimates that occur when no adjustments are made for the missing data.

The following example uses existing LSYPE weights to illustrate the effect of compensating for missing data on analyses of student attainment. Consider the estimation of the percentage of young people achieving the equivalent of five General Certificates of Secondary Education (GCSEs) at grades A*-C (level 2). Table 2-1 shows three different estimates of this statistic. The first is based on young people in households who responded to Wave 1 of the LSYPE and uses the final Wave 1 cross-sectional weight, W1FINWT. The second is based on young people in all households who responded to Wave 3 of the survey and again uses the Wave 1 weight. The third estimate is based on young people in all households who responded to Wave 3 and uses the final Wave 3 cross-sectional weight, W3FINWT. In all three cases, the set of young people is further restricted to those for whom a non-missing level 2 attainment variable is available.

Table 2-1 Estimated percentage of students achieving the equivalent of five GCSEs at grades A*-C, by sample and survey weight

Sample	Sample size	Survey weight	Estimate (%)	Standard error (%)
Wave 1 responding households	15,330	W1FINWT	58.9	0.84
Wave 3 responding households	12,295	W3FINWT	59.1	0.88
Wave 3 responding households	12,295	W1FINWT	62.6	0.85

Using the LSYPE data, the best estimate of the percentage of students achieving level 2 is the one based on Wave 1 responding households: 58.9 percent. The estimate based on Wave 3 responding households and the Wave 3 final weight, 59.1 percent, is not appreciably different, which demonstrates that the weighting adjustments used to compensate for attrition non-response worked well. However, the last row of the table shows that the percentage of students achieving level 2 would have been over-estimated if the sample had been restricted to those households that continued to participate in the survey at Wave 3 without adjustments being made for attrition non-response (i.e., no adjustments are made to the Wave 1 final weight).

Continuing to look at the percentage of students who achieved the equivalent of five GCSEs at grades A*-C, Table 2-2 shows estimates based on two subgroups of the Wave 1 responding households - those in which both the sampled young person and main parent responded to the survey and those in which either did not - using the Wave 1 final weight.

Table 2-2 Estimated percentage of students achieving the equivalent of five GCSEs at grades A*-C, by young person and main parent response status

Sample	Sample size	Estimate (%)	Standard error (%)
Wave 1 households with both a young person respondent and a main parent respondent	14,369	59.7	0.85
Wave 1 responding households with no young person respondent or no main parent respondent	961	44.3	2.31

This example highlights the issue of partial non-response, whereby responses are obtained to some but not all components of the survey. When applied to the subset of Wave 1 responding households in which the young person or main parent did not complete the survey, the weight, W1FINWT, produces a significant under-estimate of the percentage of students achieving level 2 proficiency.

The use of weights for estimating population parameters like means and proportions is generally well-accepted. However, their use for regression analyses is more debatable. When they are used, the resultant estimates of the regression parameters are estimates of the parameters that would be obtained if the regression analysis were applied to the full population. In the case of a mis-specified model, the estimates relate to that mis-specified model. The weighted estimates provide some degree of protection against mis-specification in the sense that they produce the best-fitting model of the chosen form for the given population. If it were possible to establish a “correct” model, then the weighted and unweighted estimates would estimate the same regression parameters, and the unweighted estimates would have lower variances. However,

some degree of mis-specification is inevitable. Korn and Graubard (1995; 1999, Chapter 4) discuss some examples where weighted and unweighted regression analyses yield markedly different parameter estimates.

Methods that are sometimes used for dealing with item non-response in multivariate analyses, such as multiple regression analyses, include complete case analysis and available case analysis. Complete case analysis includes only those cases that have responses for the full set of variables included in the analysis. The limitation of this approach is that when there are several variables involved, each of which may be subject to only a small amount of missing data, the number of complete cases may nevertheless be seriously reduced, resulting in large sampling errors for the estimates produced. Moreover, the complete cases are likely to be far from MCAR, resulting in biased estimates. The available case analysis uses all the available data to estimate the means, variances, and co-variances for the analysis, but it can lead to problems when the estimates based on different subsets of cases are combined in the analysis. Imputation creates a complete data set by filling in the missing responses. It thus makes full use of the reported data while also avoiding the problems with the available case analysis. In a study based on data from the U.S. component of the International Reading Literacy Study, Winglee et al. (2001) compared the results obtained from complete case, available case, and imputation-based multiple regression analyses predicting reading performance scores from such variables as gender, age, race, parents' education, and family wealth, and also a method based on the expectation maximization (EM) algorithm. They found that the methods produced similar results and concluded that the simplicity of analyses of a data set completed by imputation make it an attractive option for most analyses.

3 Imputation vs Weighting

3.1 General Considerations

With imputation, missing responses are assigned values, with all the recorded responses retained. Thus, this approach for handling missing data has the attraction of making full use of the reported data. By using responses to other items in imputing for a missing response, imputation can reduce biases in survey estimates. Estimates from an imputed data set can be produced in the same way that they would be produced from a complete data set. However, the standard errors for the estimates produced from an imputed data set using standard methods of analysis will be incorrect because the imputed values are being treated as if they are actual responses. If the amount of imputation is small, the standard errors computed using standard methods may be reasonable, but not otherwise. A variety of methods have been developed to compute standard errors correctly when imputation has been used, but to date these methods are not routine to apply and no method is entirely satisfactory.

The dominant concern with imputation is that it may affect the associations between items. In general, the association between an imputed item and another item is attenuated towards zero unless the other item is included as an auxiliary variable in the imputation scheme. Methods have been developed for incorporating all other survey items in an imputation scheme for relatively straightforward cross-sectional surveys. However, the challenges of taking into account all variables that have some degree of association with the item being imputed become extremely severe with a complex longitudinal survey like the LSYPE.

In contrast to imputation, the weighting solution for non-response incorporates information known for the non-respondents only through the few variables used in making the non-response weighting adjustments. Weighting adjustments are the standard method used for unit (total) non-response, when no survey data are collected for the non-respondents. All that is known about unit non-respondents consists of the usually very limited amount of data available on the sampling frame. In this case, these data can be incorporated in the weighting adjustments. No survey data are lost by dropping unit non-respondents from the analysis file. Weighting adjustments and imputation are inter-related. Weighting adjustments can be viewed as imputation of complete records from matched sets of respondents. As such, weighting adjustments retain all of the relationships present in the matched set of respondent records.

We recommend the use of imputation to handle item non-response and weighting adjustments to handle unit non-response. The choice between imputation and weighting for component non-response within waves and across waves is less straightforward and is discussed in the next section.

3.2 Choosing Between Weighting and Mass Imputation for Partial Non-response

Partial non-response - such as a missing interview with one of the parents - creates blocks of missing items in the longitudinal data set. When imputation is used to compensate for partial non-response, it serves to retain all the cases with some completed components by filling in holes in the rest of the data. In general, the challenge with this mass imputation is to maintain all the important associations in the data set. These include (1) associations among the imputed items in the block of missing items; (2) cross-sectional associations between the imputed items

and reported items in the same wave; (3) associations between the imputed items in the given wave and the corresponding items in other waves, particularly the same items at neighbouring waves; and (4) associations between the imputed items in the given wave and other items in neighbouring waves. In practice it is not possible to maintain all these associations with a large and highly complex data set like that of the LSYPE, and some relaxation of these goals would be needed. It should also be noted that while imputation may appear to maintain the full sample size, the effective sample size of an analysis using imputed data is smaller because the imputed items have not been directly observed.

If mass imputation were applied for partial non-response in the LSYPE, it would probably be necessary to impute missing data when they first appear and then to re-impute them when reported data become available in subsequent waves. For example, when Wave 4 data are newly collected, mass imputation may be applied for a missing Wave 4 young person questionnaire using data from Wave 3 in the imputation scheme. This approach would permit an imputed data set to be made available to users at that time for analyses of Waves 1 to 4. When Wave 5 data become available, it would probably be necessary to re-impute the missing Wave 4 data to take account of the information from both Waves 3 and 5 in order to avoid attenuation of the associations between Wave 4 and Wave 5 responses.

Mass imputation is sometimes used to compensate for partial non-response. It is, for example, used in the U.S. Survey of Income and Program Participation (SIPP), a panel survey of limited duration in which sampled households are interviewed at intervals of four months. Imputation is used in the SIPP for cases of wave non-response in which one or two consecutive waves of data are missing, provided that the missing waves are bounded by reported waves (Kalton et al., 1998). Carry-over imputation methods are used for many items. For example, if a household reports receiving Social Security income in each of the bounding waves, the household is imputed as receiving that income in the missing waves. If it reports receipt in one of the bounding waves but not the other, the month in which receipt ended or began is imputed by a random method. This approach is attractive when a key objective for the survey is to aggregate amounts across waves, such as to produce annual Social Security income, but it is less attractive for other forms of analysis.

As noted in the previous section, the advantage of weighting is that it maintains the associations between all the items in the respondent data set. When applied with partial non-response, however, weighting inevitably loses some of the partial information that is available. As an extreme example, if a single weighting adjustment were to be used to compensate for non-response to any of the data collection components, only the subset of cases with all the components completed would be included in the analysis file. The weighting adjustment would compensate for all the other cases. Thus, this procedure would lead to a considerable reduction in sample size. For example, analyses of the young person data alone would exclude the young person data for cases where the parent data were missing but the young person data were not. The use of multiple sets of weights for different combinations of components can often address this problem to an acceptable extent.

Weighting adjustments are widely used to compensate for wave non-response in longitudinal surveys. To illustrate the approach, consider first a simple longitudinal study with four waves. If attempts are made to collect data from all the originally sampled units at each wave, there are 16 different patterns of response/non-response across waves (Kalton, 1986). For example, some respondents will respond to all waves, some will respond only to the first wave, and some will respond to the first, second, and fourth waves. Suppose that an analyst wants to use the data from Waves 1 and 4 only. The analyst would then want to include all those cases with data

reported for both these waves. With the weighting adjustment approach, a weighting adjustment would be developed to compensate for all the cases that were non-respondents to one or both of these waves. (If no attempts are made to collect data at later waves from first-wave non-respondents, there are only eight different patterns of response/non-response; there are only four sets of weights if no attempts are made to re-contact previous wave non-respondents, as is the case with the LSYPE households thus far.)

This basic approach can be extended to deal with cases of partial non-response in the LSYPE, including missing waves and missing components. However, because partial non-response may arise from any of several components at each wave, the weighting approach could lead rapidly to a proliferation of sets of weights that are applicable for analyses of different combinations of data. If separate weights were produced for each observed combination, the large number of sets of weights would be problematic both for the work required to develop the weights and for users to be able to choose the correct set for their particular analysis. For these reasons, a compromise solution is often developed: Specific sets of weights that include all cases with the required data are not created for every possible analysis, but the compromise attempts to guarantee that a suitable weight will be available to include nearly all eligible cases for any given analysis, with special attention being paid to retaining sample size for critical analyses that span important parts of the analytic space. To apply this approach, we have examined the patterns of partial non-response by component and wave. The decisions to be made for this compromise involved an assessment of (1) the important combinations of waves and components for analyses and (2) how many cases would be lost if weighting adjustments were produced only for a combination of waves and components that included more waves and/or combinations than needed for a given analysis.

We recommend the use of the weighting adjustment approach for missing components in the LSYPE. The next chapter examines the patterns of component non-response across the first three waves of the survey in order to develop recommendations for the sets of weights to be computed.

4 Weighting for Missing Waves and Components

4.1 Component Combinations of Significant Analytic Interest

The first step in developing recommendations for the sets of weights to use to compensate for component non-response in the first three waves of the LSYPE was to identify the various components and to establish which combinations of components are of major analytic interest. In the first two waves of the survey, data were collected from the young person, the main parent, and a second parent if there was one; there was also a history component. In Wave 3, data were collected only from the young person and the main parent.

Initial investigations into the nature of missing data in the LSYPE examined the interviewer-administered and CASI sections of the young person and main parent interviews as separate components. However, it became clear from user consultation that the young person main interview and CASI section are generally regarded as one survey component. Also, even though there is some likelihood of analyses involving parent data only from the main parent interview (and not from the main parent CASI section), the two main parent components are often viewed as a whole. For example, there is much interest in the influences of academic self-concept, attitudes and aspirations with regard to schooling, bullying, engagement in positive activities, and non-cognitive outcomes (such as truancy, drug taking, and smoking). Much of these data come from the young person and main parent CASI sections, reinforcing the view that these components are an integral part of the survey and should be considered in conjunction with the corresponding main interviews. The decision was therefore made not to treat the interviewer-administered and CASI sections as separate components, but rather to make sure that sufficient data were collected from each section for the component to be classified as a response (see Section 4.2).

Little analytic interest was expressed in data from the second parent interview (with the exception of income, educational qualifications, and occupation), and it seems unlikely that these data would be of interest in the absence of data from the main parent. Various aspects of history data were collected through different questionnaire components during Waves 1 and 2 of the LSYPE. For example, items about the young person's birth weight, health, and siblings were asked during the history main interview, and some of these items were asked only of the natural mother. Items about the young person's school history were asked during the Wave 1 history main interview and during the Wave 2 main parent main interview. Items about relationship history were asked during the history CASI section and only of resident natural parents. After consultation with the DCSF, we decided to focus on the history CASI component and only in the context of longitudinal analyses. Since this component was administered at Wave 2 only if responses had not been provided to it at Wave 1, we determined that there was a response to the history component if relationship data were provided at either of the first two waves (see Section 4.2 for details).

There is considerable potential research interest in combining data from the young person, main parent, second parent, and/or history components with data from the LSYPE household grid or from the PLASC or National Pupil Database. However, the latter data sources are excluded from the tabulations in this chapter for several reasons. First, both presentation and comprehension of the missing data patterns are made considerably more difficult when additional components are introduced. Second, previous analyses of the extent of non-response to the household grid component and of missing GCSE scores in the National Pupil Database revealed low levels of

missing data. Finally, the main interest in linking data from the LSYPE with the PLASC appears to be methodological and of secondary concern.

4.2 What Constitutes Component Response?

Detailed codes are used in the LSYPE data set to describe the nature of non-response to each item. There are variables in each of the "main" wave files that indicate whether or not the respective respondent accepted the CASI section of the young person, main parent, and history components of the questionnaire. In addition, the longitudinal index file contains variables that indicate whether or not the respective respondent commenced the main interview components of the young person, main parent, second parent, and history sections. However, some respondents who initiated participation in a particular component of the questionnaire may have given up partway through, and others may simply have refused to respond to a significant proportion of the questions. This raises the question: How much of an individual's data can be missing within a component before we would declare the unit response status for that particular component to be "non-response"?

A second issue relates to inconsistencies in the data set with respect to the combinations of missing data values that appear in some records. For example, the variable that indicates whether or not the young person main interview was initiated may have a missing value code because the young person is not in the household grid for a particular wave, yet he/she may have completed this section of the questionnaire. For this reason, it was also necessary to determine for any individual if components that appeared to be "missing" actually constituted a unit response.

For each component of the LSYPE data, we attempted to deal with these issues by computing the percentage of the component's items that are missing for each individual. Note that "inapplicable" data values, such as those due to questionnaire routing, were treated as a valid response. All variables in the respective wave's main file that resulted directly from an interview question in a given component were considered. For example, an item with multiple possible responses resulted in multiple variables in the main file - each corresponding to a possible answer category - and each of these variables was examined for non-response. However, derived variables (whose labels generally contain the letters "DV") were not counted in the determination of the percentage missing. In most cases, among those who initiated response to a given component, the percentage of missing data was not large.

For these reasons, and to reduce the considerable complexity of the possible missing data patterns, we classified a young person as a respondent at a given wave only if he/she responded to at least 25 percent of the items in the young person main interview *and* at least 25 percent of the items in the young person CASI section. The same rules were applied to the two main parent components to determine the response status of the main parent at a given wave. For the history component, the criterion was responses to at least 25 percent of the items in the relationship CASI section at either Wave 1 or Wave 2. The CASI data collection sections were skipped for a non-trivial number of respondents who required an interpreter during the main component of their interview. These individuals were classified as "responding" to the CASI component under the assumption that inference would be made to the population of English speakers.

Even though the rules we have adopted did not result in the exclusion of many cases with incomplete data, they were necessarily arbitrary. There are a number of possible ways to determine whether or not the available data are sufficient to constitute a response to any given component. Consequently, the sample sizes quoted in this report should not be considered definitive.

Using the response definitions described above, Section 4.3 examines component non-response in Wave 1 of the LSYPE and recommends selected sets of cross-sectional weights to be developed for these data. The subsequent two sections extend this treatment to longitudinal patterns of missing data across the first three waves of the survey. Section 4.6 combines these findings and concludes with a summary recommendation of weighting adjustments for the LSYPE. Throughout this discussion it is important to note that the trade-off between maximizing analytic sample sizes and developing a manageable number of sets of weights necessarily involves a degree of subjectivity.

4.3 Missing Data in Wave 1

Table 4-1 shows the responding sample size for each of the four core components: Wave 1 young person, main parent, and second parent and Wave 1 or 2 relationship history. Not surprisingly, the second parent interview contains the highest level of missing data.

Table 4-1 Unit response to the four core components at Wave 1

Component	Response	Non-response	Total	Conditional response rate (%) *
Young person	15,298	472	15,770	97.0
Main parent	15,157	613	15,770	96.1
Second parent	14,288	1,482	15,770	90.6
History	14,740	1,030	15,770	93.5

* This response rate is conditional on response to at least one Wave 1 component.

Ideally, data users would want to base their research on all the cases that responded to the set of components involved in their analyses. As noted in Section 4.1, we decided to exclude consideration of the history component for cross-sectional analyses on the grounds that its data are of limited utility for such analyses. Table 4-2 presents the full range of possible sets of the other components in which analysts might be interested, along with the responding sample size at Wave 1.

Table 4-2 Analytic combinations of data from Wave 1

Wave 1 data required		Sample size
Young person only		15,298 *
Main parent only		15,157
Young person and main parent only		14,763 *
Second parent only		14,288
Young person and second parent only		13,941
Main parent and second parent only		13,832
Young person, main parent, and second parent		13,552 *

* Recommended sets of weights.

Since all the combinations in Table 4-2 are of potential interest, this would imply computing a separate set of weights for each of the seven rows of the table. However, while in principle it would be possible to compute all these sets of weights, such an approach is unattractive both because of the amount of work involved in developing the weights and because of the complexity for analysts in choosing the appropriate set of weights.

A number of possible sets of weights (second parent only, etc.) can be eliminated at the outset because of a lack of perceived utility. Among the remaining options, it is necessary to compromise between practicality and sample size. For Wave 1 of the survey, we recommend the following three sets of weights (as indicated with an asterisk in Table 4-2): one set of weights for the 15,298 households with responses to the young person component at Wave 1; a second set of weights for the 14,763 households with responses to the young person and main parent components at Wave 1; and a third set of weights for the 13,552 households with responses to all three core components (i.e., to the young person, main parent, and second parent components) at Wave 1. A fourth set of possible weights would be for the 15,157 households with responses to the main parent component at Wave 1 - information that may be of interest in connection with young person academic data from the National Pupil Database - but we have chosen not to include this set in our recommendation. Analysts of parent data would instead use the weights for the 14,763 cases with responses to both the young person and main parent components. In general, all analyses could be conducted with a single set of weights based on the last row of Table 4-2, but this would result in a considerable loss of sample size for certain data combinations.

4.4 Missing Data in Waves 1 and 2

The sets of weights recommended in Section 4.3 lead naturally to consideration of certain analytic combinations of data from Waves 1 and 2. For instance, Table 4-3 shows the responding sample sizes for selected combinations of Wave 1 data together with the young person data at Wave 2 (see column 2); the young person and main parent components at Wave 2 (see column 3); and all three core components (young person, main parent, and second parent) at Wave 2 (see column 4). It is assumed that interest in longitudinal analyses of parent data only is limited, and therefore such combinations are not considered in this table.

Table 4-3 Sample sizes for selected combinations of Wave 1 and Wave 2 data

Wave 1 data required	Wave 2 young person data	Wave 2 young person and main parent data	Wave 2 young person, main parent, and second parent data
None	13,239 *	12,852 *	11,722 *
Young person only	12,993 *	12,621 *	11,517
Main parent only	12,797	12,488	11,407
Young person and history only	12,550	12,285	11,269
Young person and main parent only	12,598	12,302	11,244
Young person, main parent, and history only	12,267 *	12,028	11,043
Young person, main parent, second parent, and history	11,418	11,210	10,582 *

* Recommended sets of weights.

Based on the first row of Table 4-3, three sets of cross-sectional weights could be computed for Wave 2 of the LSYPE, corresponding to the three main sets of weights recommended for Wave 1. These sets of weights would be for (1) the 13,239 households with responses to the young person component at Wave 2; (2) the 12,852 households with responses to the young person and main parent components at Wave 2; and (3) the 11,722 households with responses to all three core components (i.e., to the young person, main parent, and second parent components) at Wave 2.

In choosing among the remaining data combinations shown in Table 4-3, it is necessary to suffer some loss of analytic sample size to reduce the recommended sets of weights to a manageable number. Under the assumption that there is likely to be a high level of interest in analyses of the Wave 1 and 2 young person and main parent data, we recommend a set of longitudinal weights for each of the following combinations: (1) the 12,993 households with responses to the young person component at Waves 1 and 2; (2) the 12,267 households with responses to the young person and main parent components at Wave 1, the history component (at Wave 1 or 2), and the young person component at Wave 2; and (3) the 12,621 households with responses to the young person component at Wave 1 and the young person and main parent components at Wave 2. Finally, we recommend a “catch all” set of Wave 1-Wave 2 longitudinal weights to represent the 10,582 households with responses to the young person and the main and second parent components at Waves 1 and 2 and the history component at either Wave 1 or Wave 2. This last set of weights provides analysts of data from Waves 1 and 2 with a default weighting option that can be applied for analyses of any combination of components across the two waves.

4.5 Missing Data in Waves 1, 2, and 3

The number of possible analytic combinations of components involving data from Waves 1, 2, and/or 3 is very large. It is therefore necessary to restrict consideration of weighting options to those sets of weights that are most likely to be of interest to researchers. Table 4-4 shows the responding sample sizes for selected combinations of data from Wave 1 and / or 2 together with the young person data at Wave 3 (see column 2) and the young person and main parent components at Wave 3 (see column 3). (We do not consider the second parent component

separately at Wave 3 because the second parent interview was conducted jointly with the main parent interview in most households.)

Table 4-4 Sample sizes for selected combinations of data from Waves 1, 2, and 3

Waves 1 and 2 data required	Wave 3 young person data	Wave 3 young person and main parent data
None	12,243 *	11,893 *
Wave 1 young person only	12,026	11,690
Wave 2 young person only	12,064	11,724
Waves 1 and 2 young person only	11,866 *	11,539
Wave 1 main parent only	11,847	11,553
Wave 2 main parent only	11,911	11,643
Waves 1 and 2 main parent only	11,585	11,344
Wave 1 young person and main parent only	11,672	11,389
Wave 2 young person and main parent only	11,757	11,493
Wave 1 young person and main parent, and Wave 2 young person only	11,520	11,244
Waves 1 and 2 young person and main parent only	11,287	11,057
Wave 1 young person, main parent, and history only	11,385	11,129
Wave 2 young person, main parent, and history only	11,448	11,203
Wave 1 young person and main parent, Wave 2 young person, and history only	11,241 *	10,991
Waves 1 and 2 young person and main parent, and history only	11,051	10,834
Wave 1 young person, main parent, second parent, and history only	10,612	10,381
Wave 1 young person, main parent, and second parent, Wave 2 young person, and history only	10,483	10,257 *

* Recommended sets of weights.

Corresponding to the sample sizes in the first data row of Table 4-4, we recommend that two sets of cross-sectional weights be computed for Wave 3 of the LSYPE. One of these sets of weights would represent the 12,243 households with responses to the young person component at Wave 3, and the other would represent the 11,893 households with responses to the young person and main parent components at Wave 3. With regard to Wave 1-Wave 3 longitudinal weights, we recommend a set of weights to represent each of the following combinations: (1) the 11,866 households with responses to the young person component at Waves 1, 2, and 3; (2) the 11,241 households with responses to the young person and main parent components at Wave 1, the young person component at Wave 2, the history component (at Wave 1 or 2), and the young person component at Wave 3; and (3) the 10,257 households with responses to the young person and the main and second parent components at Wave 1, the young person component at Wave 2, the history component (at Wave 1 or 2), and the young person and main parent components at Wave 3. The first set of longitudinal weights described above [i.e., (1)] is a natural choice in that it can be used to analyse young people who responded in each of the first three waves of the LSYPE. The second set of longitudinal weights (2) is intended to facilitate

longitudinal analyses of the young person data that require some information about the main parent. This set of weights applies only to those young people who responded at all three waves and who are from households for which history data are available; however, the decrease in sample size compared to similar alternative data combinations (e.g., those for which young people data are available only for Waves 1 and 3) is not appreciable. The final set of weights provides analysts of data from the first three waves with a “catch all” weighting option that can be used for combinations of components across waves not covered by the preceding sets of weights.

4.6 Recommendations for Weighting

In the previous sections of this chapter, we examined the patterns of missing data among the first three waves of the LSYPE and made some preliminary recommendations for sets of weights to be computed. To summarise these findings, Table 4-5 provides an initial list of recommended sets of weights.

Table 4-5 Recommended sets of weights for Waves 1 through 3 of the LSYPE

Weight description	History data required	Wave 1 data required	Wave 2 data required	Wave 3 data required	Sample size
CROSS-SECTIONAL					
Wave 1		Young person only			15,298
Wave 1		Young person and main parent only			14,763
Wave 1		Young person, main parent, and second parent			13,552
Wave 2			Young person only		13,239
Wave 2			Young person and main parent only		12,852
Wave 2			Young person, main parent, and second parent		11,722
Wave 3				Young person only	12,243
Wave 3				Young person and main parent	11,893
LONGITUDINAL					
Waves 1-2		Young person only	Young person only		12,993
Waves 1-2	Yes	Young person and main parent only	Young person only		12,267
Waves 1-2		Young person only	Young person and main parent only		12,621
Waves 1-2	Yes	Young person, main parent, and second parent	Young person, main parent, and second parent		10,582
Waves 1-3		Young person only	Young person only	Young person only	11,866
Waves 1-3	Yes	Young person and main parent only	Young person only	Young person only	11,241
Waves 1-3	Yes	Young person, main parent, and second parent	Young person only	Young person and main parent only	10,257

Table 4-5 lists a total of 15 sets of weights, 8 of which are intended for cross-sectional analyses and 7 of which are selected for longitudinal analyses. While it is not unusual for a large number of sets of weights to be required for a panel survey as complex as the LSYPE, it would be beneficial to reduce the number of sets of weights to be produced for the first three waves of the LSYPE, both for user convenience and to reduce the effort required to develop the weighting adjustments.

In conclusion, we should emphasise the subjectivity of our choices. They are based on an assessment of the most useful subsets of components for analyses. Others may, however, prefer alternative subsets.

5 Imputation for Item Non-response

The use of weighting adjustments such as those recommended in Chapter 4 would address wave and component non-response in the LSYPE for most research interests. However, item non-response amongst otherwise complete components remains an issue. Analysts could use a method such as complete case analysis to deal with item non-response, but this can seriously reduce the available sample size and result in biased estimates. While it would be convenient for analysts to have a data set in which missing values have been filled in, careful imputing for all missing data items in the LSYPE would be an enormous task because of the size and complexity of the survey. A more feasible approach is to impute for items that are likely to be key for many analyses.

When imputation is used for a key survey item, care should be taken in the imputation scheme to include the major auxiliary variables associated with that item. With an ongoing longitudinal study, this approach needs to deal with the issue that some of the main auxiliary variables are collected in other waves. It is very important to maintain the associations between the responses to the same item measured at different waves, particularly the adjacent waves. However, item imputation faces a timing challenge. Responses from the previous wave can be incorporated in imputing for an item at the current wave, and an analytic file can be released without delay. However, it may be desirable to revise the imputations to take account of response at the subsequent wave (as has been done in the British Household Panel Survey).

Hot deck imputation is used for many surveys, and this method would likely serve well for some of the LSYPE items. In general, this procedure consists of randomly matching observations within cells defined by auxiliary variables, in the search for a donor from which to obtain an imputed value. For many items subject to imputation, the amount of missing data will be small, and a simple hot deck based on demographic auxiliary variables should suffice. However, regression-based methods, combined with hot deck imputations of residuals, may be advisable for key items that have many important predictor variables, such as household income. With this approach, a regression model for income is built in terms of items such as educational qualifications, gender, and age, and a predicted value is generated from the model for each case in the data set. Hot deck imputation is then applied to the residuals from the regression model, and the residual belonging to the donor case is added to the predicted income value for the case requiring imputation.

The use of Bayesian parametric algorithms for data imputation - for example, IVEware (Raghunathan, Solenberger, and Van Hoewyk, 2002) and MICE (van Buuren and Oudshoorn, 2000) - has grown in recent years. Several statistical software packages offer built-in or add-on imputation modules that implement regression-based Bayesian methods. The basic idea is to draw imputed values from a posterior predictive distribution specified by a regression model, usually with a flat or non-informative prior distribution for the regression parameters. While this approach should do better than traditional hot deck imputation at preserving multivariate structure, it too has its disadvantages. For example, Bayesian methods are often heavily reliant on normality assumptions and are not designed to cope well with unusually shaped distributions, such as heaping of reported income at round thousands. Their ability to produce imputed data that adhere to questionnaire skip patterns is often limited, which can be problematic when working with survey data. Also, despite advances in computing power, substantial expense can be involved in monitoring the convergence of Monte Carlo Markov chains.

Westat has developed a regression-based program (AutoImpute) for imputing for "Swiss cheese" patterns of item non-response in an efficient manner that incorporates auxiliary items that may also be subject to missingness. Piesse, Judkins, and Fan (2005) describe the statistical methodology used by AutoImpute, which blends ideas from Gibbs sampling, data mining, predictive mean matching, and hot deck imputation. In brief, a simple hot deck is used to initialise the imputation process so that all items can then be used to predict all other items defined on comparable sets of cases, regardless of the complexity of the overall missing data pattern. Each item is then re-imputed in turn. After all the items have been re-imputed once, convergence is assessed through the R-squared statistics from the prediction models. Judkins et al. (2007) describe recent enhancements and an evaluation of the software.

The following important topic areas in the LSYPE data set have been identified as the primary candidates for imputation:

- Household income;
- Mother's, father's, and highest parental qualifications (where the parent resides with the young person);
- Family's National Statistics Socio-economic Classification (NS-SEC);
- Attitudes and aspirations towards schooling; and
- Bullying.

Imputation of household income is challenging due to high missing rates (in excess of 30 percent at Wave 1) and the concentration of non-response in particular subsets of the cohort, such as Bangladeshi, Pakistani, and other South Asian ethnic groups. Aside from the missing data issue, a considerable problem is that the LSYPE items measuring income changed across waves. In the Wave 1 main parent interview, the main parent was asked to provide the total income of both parents from all sources and to provide information about receipt of benefits by either parent. The main and second parents gave separate reports on their hourly wages and salaries. In Wave 2, the main and second parents gave separate reports on earnings (hourly wage and salary) and benefits (pensions, child or disability allowances, etc.). In Wave 3, the main parent and second parent interviews were usually conducted jointly; each parent was asked to provide the total income of both parents from all sources (and to identify those sources). In this specific situation, a first step towards imputing household income might be to compute one or more derived variables that represent comparable income measures across survey waves.[1] These derived income variables could then be imputed following the general imputation guidelines outlined above (for example, using information from non-missing data values in neighbouring waves and taking account of strongly associated items).

[1] For example, the labels on the variables W1INC1EST, W2INC1ESTMP, and W3INCESTMP in the main files for Waves 1, 2, and 3, respectively, suggest that each is a measure of the combined gross income from all sources of the main and second parents. However, W2INC1ESTMP does not appear to be a derived variable, and the corresponding Wave 2 question asks the main parent about "your gross pay" - that is, it does not appear to include income from non-work-related sources or from the second parent. Similarly, the variables W1GRSSYRHHBANDS and W2GRSSYRHHBANDS appear to represent household income bands based on Wave 1 and Wave 2 salaries, respectively, but there does not appear to be an equivalent variable on the Wave 3 main file.

During Waves 1 and 2 of the LSYPE, attempts were made to gather information on the educational qualifications of the main and second parents. Item non-response rates for mother's qualifications are around 6 percent in Wave 1 and 2 percent in Wave 2; however, the rates for father's educational qualifications are at least twice as high in both waves. The primary social classification used in the United Kingdom, the NS-SEC, is occupation-based. Detailed information on the NS-SEC was collected only in Waves 1 and 2. The item missing rates for the derived variables representing the family's NS-SEC class at Waves 1 and 2 (W1NSSECFAM and W2NSSECFAM, respectively) are approximately 12 percent.

To take advantage of and maintain the strong associations between items related to income, parental qualifications, and NS-SEC, we recommend that they be imputed collectively. The use of an imputation method that iterates or cycles through the data set until acceptable convergence criteria have been reached (such as AutoImpute) may be advantageous for this cluster of data items.

Data on topics such as attitudes towards school and bullying are collected through batteries of questions in the LSYPE, each of which may be subject to item non-response. For example, there are a number of items about different types of bullying and the frequency with which each was experienced. An issue with data collected in this way is that non-response to any one of a number of items may lead to a missing value for a derived composite variable. This can lead rapidly to a decrease in the number of cases with non-missing values for the composite variable of interest. Consider the sequence of items about bullying in the Wave 1 young person CASI component. There are five questions asking the young person whether or not he/she experienced a particular type of bullying in the last 12 months. Corresponding to each affirmative response, the young person is then asked to report how often that form of bullying occurred. Missing data can therefore arise from either the initial question about occurrence or the follow-up question about frequency (if applicable) through a refusal or "don't know" response. Among those young people who responded to the CASI component at Wave 1, Table 5-1 shows the rate of missingness among the frequency-of-bullying items (where zero frequency has been logically imputed for those who responded "no" to the lead question).

Table 5-1 Percentage of missing data among Wave 1 bullying items

Wave 1 bullying item	Missing data (%)
How often upset by name-calling in last 12 months	6.7
How often excluded from a group of friends in last 12 months	5.5
How often been made to hand over money or possessions in last 12 months	2.4
How often threatened with violence by students in last 12 months	4.2
How often experienced violence from students in last 12 months	3.8

The item non-response rates in Table 5-1 are modest when the bullying items are considered individually. However, for a derived variable based on all five of these items, the missing rate would be determined by the non-response pattern across the items. Table 5-2 shows the percentage of young people with missing data for none of the bullying items and for one or more of the five items, among those who responded to the CASI component.

Table 5-2 Distribution of the number of bullying items with missing frequency data at Wave 1

Count of bullying items with missing frequency data	Percentage of young people (%)
0	84.8
1	10.4
2	2.9
3	1.2
4	0.4
5	0.2

From the first row of Table 5-2, it is evident that a composite variable based on frequency of bullying may be subject to a missing rate as high as 15 percent (if the composite requires responses to all five items). This is considerably higher than the non-response rate for any of the individual items. It is also noteworthy that the 85 percent of young people who responded to all five items differ from the 15 percent who did not. This is illustrated in Table 5-3 in terms of their attainment of level 2 proficiency, with estimates computed using the Wave 1 final weight. Young people who responded to all the bullying items at Wave 1 were more likely to achieve the equivalent of five GCSEs at grades A*-C.

Table 5-3 Estimated percentage of students achieving the equivalent of five GCSEs at grades A*-C, by response to Wave 1 bullying items

Sample	Sample size	Estimate (%)	Standard error (%)
Young people who responded to all five bullying items at Wave 1	12,569	61.8	0.83
Young people who responded to the Wave 1 CASI component but did not answer all five bullying items	2,220	48.3	1.42

The imputation approach that allows for maximum flexibility when defining and computing composite variables is to impute for each of the possible constituent items. Another strategy would be to impute the composite variables directly; however, this would not take account of reported constituent items when others in the same cluster or sequence are missing. Once the candidate variables for imputation have been identified, these should be imputed by borrowing information from non-missing values of the same variables in neighbouring waves and completed data items within the same questionnaire sequence where possible, as well as taking account of demographic characteristics such as gender and ethnicity. If all items in a given sequence are subject to non-response, data should be imputed using a hot deck approach or one of the software alternatives mentioned earlier in this chapter.

An issue relating to attitudinal data items is how to treat “don’t know” responses. These may be deemed valid answers to some questions, but not to others. However, when the alternative responses are on a clearly ordinal scale, it is often difficult to determine where within the scale a “don’t know” response should lie. For these reasons, imputation for “don’t know” responses to key attitudinal items should be considered on an individual basis.

Measures that are reasonably stable over time (such as parental qualifications) and those for which the best information has likely already been collected in the first three waves of the survey (such as income) may be imputed once. For items such as bullying, there may be a need for initial imputation as the data for a given wave are being processed, followed by re-imputation as later survey data become available. However, efforts to improve upon previously imputed values must be weighed against the cost implications.

It should also be noted that imputation is a general-purpose strategy for handling missing data. It is intended to compensate for item non-response in a way that should meet the needs of most users, but some may prefer alternative methods for handling the missing data. Therefore, it is a standard best practice to “flag” imputed values in the data set so that users can identify them. The flags enable users to discard the imputed values for an analysis where they prefer to deal with the missing responses in some other tailor-made way specific for that analysis.

The topic areas recommended for imputation above are likely to be key for many analyses. In future, items other than those identified in this chapter may also be considered for imputation based on the amount of missing data and the analytic significance of the item.

6 Guiding Users on Statistical Analyses

Estimation of analytic statistics of interest from a data set in which weighting adjustments and imputation have been used is straightforward. However, variance estimation is more complicated. Treating the weighting adjustments as fixed and the imputed values as if they were reported values leads to under-estimation of the variances of the survey estimates. One approach to variance estimation is to use a replication methodology (either jackknife or balanced repeated replications), with the replicate weights being computed separately in each replicate in order to account for the effects of the weighting adjustments on the precision of the survey estimates. These weights can be used for variance estimation in WesVar, SUDAAN, or other packages that can apply replication methods. Handling weighting adjustments for complex sample designs with variance estimation based on Taylor series approximations is less straightforward. However, both our experience and published research (Valliant, 2004) suggest that users of SPSS/Complex Samples and similar software based on Taylor series approximations are likely to obtain standard errors acceptably close to the replication equivalents.

For many items subject to imputation, the amount of imputation will be small, and it will be reasonable to ignore the fact that a few responses were imputed. There is as yet no perfect solution for taking account of imputed values in variance estimation when the survey items are subject to substantial amounts of imputation and the sample is based on a complex sample design that requires the use of weights in the analysis. Multiple imputation can work in some cases, but it is not always correct (Kim et al., 2006). Other techniques like those of Särndal (1992), Rao and Shao (1992), Haziza and Rao (2006), and Kim and Fuller (2004) can be useful in some simple cases but are limited in applicability. At this stage of the research in this area, it would seem best to run some simple analyses based on one of these techniques and then to attempt to generate some approximate models from which to predict the magnitude of variance increase resulting from the presence of a substantial number of imputed values. These models could be provided as part of the LSYPE user's guide, with instructions on how they can be used for different forms of analysis.

The major question that analysts face with the use of survey weights is which set of weights to apply for a particular analysis. The LSYPE user's guide can help analysts answer this question. In essence, the choice of a set of weights involves determining which of the computed sets of weights that cover the survey components required for the analysis has the largest sample size. That set of weights can be readily selected from a table like Table 4-5. Including examples of the appropriate choices for selected analyses in the user's guide can aid analysts in understanding this process.

As has been noted earlier, a major challenge with imputation is to maintain the associations with all the other variables in the survey data set. Despite all the efforts that are taken to maintain these associations, there is inevitably some slippage. In the extreme case, there is even the risk that an imputed value may be inconsistent with a combination of responses to other survey items. For this reason, it is a standard best practice to "flag" imputed values in the data set so that users can identify them. The flags enable users to discard the imputed values for an analysis where they consider that the imputations may be problematic, dealing with the missing responses in some other tailor-made way specific for that analysis. The flags also enable users to examine whether imputation may be the cause of any anomalous values. Users need to be made aware of the meaning of these flags in the survey documentation.

References

Haziza, D., and Rao, J.N.K. (2006). A non-response model approach to inference under imputation for missing survey data. *Survey Methodology*, 32, 53-64.

Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Huang, W.C. (2007). Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association* [CD-ROM], pp. 3211-3218. Alexandria, VA: American Statistical Association.

Kalton, G. (1986). Handling wave non-response in panel surveys. *Journal of Official Statistics*, 2, 303-314.

Kalton, G., Winglee, M., Rizzo, L., Jabine, T., and Levine, D. (1998). *SIPP Quality Profile 1998*. 3rd ed. Washington, D.C.: U.S. Census Bureau.

Kim, J.K., Brick, J.M., Fuller, W.A., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society, Ser. B.,* 68, 509-521.

Kim, J.K. and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika,* 91, 559-578.

Korn, E.L. and Graubard, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.

Korn, E.L. and Graubard, B.I. (1999). *Analysis of health surveys*. New York: Wiley.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Piesse, A., Judkins, D., and Fan, Z. (2005). Item imputation made easy. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3476-3479.

Raghunathan, T.E., Solenberger, P.W., and Van Hoewyk, J. (2002). *IVEware: Imputation and variance estimation software user guide*. Ann Arbor: Institute for Social Research, University of Michigan.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Särndal, C.E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology,* 18, 241-252.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

Van Buuren, S. and Oudshoorn, C.G.M. (2000). *Multivariate imputation by chained equations: MICE V1.0 user's manual*. TNO report PG/VGZ/00.038. Leiden: Netherlands Organization for Applied Scientific Research.

Winglee, M., Kalton, G., Rust, K., and Kasprzyk, D. (2001). Handling item non-response in the U.S. component of the IEA Reading Literacy Study. *Journal of Educational and Behavioral Statistics*, 26(3), 343-359.

Ref: DCSF-RW086

ISBN: 978 1 84775 337 3

www.dcsf.gov.uk/research

Published by the Department for
Children, Schools and Families

department for
children, schools and families