The use of linked administrative data to tackle non response and attrition in longitudinal studies

Andrew Ledger & James Halse

Department for Children, Schools & Families (UK)
Andrew.Ledger@dcsf.gsi.gov.uk
James.Halse@dcsf.gsi.gov.uk

Abstract
Most social surveys encounter problems through non response. In longitudinal studies, this problem is compounded by attrition between waves and the combined non response across the course of a study can lead to significant bias and inflation of errors. It is often possible to deal with the effects of attrition in later waves through use of data collected in earlier waves before a subject dropped out of the study. Techniques like CHAID analysis or logistic regression can then be used to weight the data back to be representative of the data from the first wave. However, most often we know nothing about those who do not respond at the first wave and this is generally where non response is highest. Although it is possible to calibrate data to take some account of this, it can be a serious problem if non response patterns contain significant bias.

This paper considers the improvements that can be made to longitudinal studies through the availability of individualised administrative records. In particular, administrative data offers information about those who do not respond to surveys. This enables a more sophisticated approach towards analysing non response and attrition and also enables better planning for future studies. With reference to two large scale longitudinal studies of young people carried out in England, this paper shows how the availability of administrative data can be used to make significant methodological improvements.

Sources of data used in this work
This paper focuses on two surveys: the Longitudinal Study of Young People in England (LSYPE) and the Youth Cohort Study (YCS). YCS, which has run for more than 20 years, has mainly been a postal survey with telephone follow up where possible for non-respondents while LSYPE has used face to face interviewing so far over 4 waves. Both studies try to track young people around the time when they finish compulsory education (YCS starts at age 16/17, LSYPE starts at 13/14) and to follow them to further education or into the labour market. YCS has just embarked on its 13th cohort while LSYPE has only had one. Both are fairly large exercises. YCS has tended to have an issued sample of around 30 thousand for the first wave, while LSYPE had an issued sample of 21 thousand at the first wave.

In addition to the survey data, in recent years, it has become possible to use the National Pupil Database (NPD); an administrative record of all pupils in maintained (not independent) schools in England. This record contains limited demographic information as well as individual level examination results and other indicators – such as whether a child is entitled to Free School Meals (FSM - a proxy for poverty) or whether a child has Special Educational Needs (SEN). The main interest as regards academic attainment is in a young person's attainment in examinations known as GCSEs. These are taken at age 15/16 and whether or not a young person attains 5 'good' GCSEs (meaning at grades A*, A, B or C) is often used as a threshold measure for success. Other academic attainment is recorded at younger ages, but is used less often.

NPD provides a reasonably complete record and can provide the basis for a sampling frame, or can be linked in to survey responses at a later stage. Around 7 per cent of young people attend independent schools and will not have a full record, although many of these will have been at a maintained school at some point and will have some details recorded. For the rest, some information is still available in the form of results for the main compulsory school age examinations (GCSEs), but little else is recorded.

Problems of non-response and attrition
Much attention is paid to response rates at the first wave of a longitudinal study. However, while this is an important measure, also important are the patterns of response and the levels of attrition between waves. YCS offers a good illustration of the difficulties faced in longitudinal studies. For many years the response rate has been falling and serious biases are evident from the data. At wave 1, there is a large discrepancy between response rates between higher and lower attainers, but this is exacerbated further as attrition affects lower attainers disproportionately. Clearly with only some 3 per cent remaining after wave 4 of those in the lowest attainment bands, results have to be treated with caution and this is especially unfortunate as these people are often of greatest policy interest.

Table 1: Cumulative response rates in YCS 12, by wave and by attainment of sample

	Wave 1	Wave 2	Wave 3	Wave 4
GCSE attainment				
None	19	9	5	3
1-4 D-G	21	10	6	3
5+ D-G	30	18	11	6
1-4 A*-C	38	23	15	9
5-7 A*-C	47	32	22	14
8+ A*-C	63	48	36	24
All	46	32	22	14

Source Youth Cohort Study, Cohort 12, waves 1-4, 2004-2007

It is reasonably clear how we might apply weighting to adjust for the problems caused by non-response. By calibrating the data so that, for example, the level of attainment of the responding sample members is equal to the level of attainment of the overall cohort, we can improve the usefulness of the data. This technique can be improved further by use of other variables. The weighting procedure for the first wave of Cohort 12 of YCS involved the construction of 3 sets of weights: selection weights; ethnic boost weights and non-response weights. The first two are readily calculable before the start of fieldwork, but clearly the final component part of the overall weights (the product of the three detailed above) depends on what happens during fieldwork. The technique used was cell weighting in which a population matrix is compared against a sample matrix and response rates calculated for each cell. Choosing the variables that define the matrix is not always straightforward, but as with the previous cohort weighting, the variables used were region, school type, gender and level of attainment.

To carry out an exercise like this, we may use administrative data for those individuals who did not respond or we may not – because we know what national figures are, the aggregated administrative data are just about good enough on their own. However, where individual level non response is especially important is in enabling us to look in detail at the characteristics of non-responders. If, early on in a survey, we can tell that the response rate from certain groups of young people is low (perhaps those of a certain ethnic background, for example), then we can ensure that special measures are taken to compensate. This might include boosting the survey for this group, attempting to convert early refusals, moving senior interviewers onto such cases and anything else appropriate. In this way we can mitigate against some of the problems about which we would otherwise find out only when we have finished the survey.

The following table shows how response rates can vary across groups, especially in a postal survey such as YCS Cohort 12, wave 1. This analysis uses administrative data and covers only young people in maintained schools in England (the full sample for YCS includes young people in independent schools and also a sample of young people from Wales – another of the constituent countries of the UK). The differences are not as marked as in table 1 since it is attainment that tends to have the strongest influence on response. However, there are still some marked differences – even by gender, we have more than a 10 percentage point difference in response at the first wave. Much of these differences can be explained by reference to the differences present by attainment. Boys tend to have lower attainment than girls and those eligible for Free School Meals tend to have lower attainment. The situation by ethnicity is less clear. Some ethnic groups have higher attainment that the white majority – in particular those from Indian and Chinese backgrounds, although here these are mixed with groups such as Pakistani and Bangladeshi who tend to have lower attainment and lower response rates.

Table 2: Cumulative response rates in YCS 12, by wave and by characteristics

	Wave 1	Wave 2	Wave 3	Wave 4
Female	52	37	27	18
Male	40	27	18	11
Non FSM	48	34	24	16
FSM	31	18	12	7
Asian	45	32	23	16
Black	31	21	14	8
Other ethnic group	42	28	20	14
Unclassified	43	30	20	11
White	47	33	23	15
Total	46	32	22	14

Source Youth Cohort Study, Cohort 12, waves 1-4, 2004-2007

Using administrative data for planning surveys

As well as early intervention, the use of administrative data is fundamental in enabling better planning for future studies. In the planning stages for YCS 13, it was possible to use the data from Cohort 12 to simulate the changing composition of the sample given the likely levels of response at each wave of the survey. An initial aim of this developmental work was to ensure a nationally representative sample by the end of the fourth wave of Cohort 13 on the assumption that Cohort 13 was to be carried out in much the same way as previous cohorts. However, based on the patterns of response that we observed in Cohort 12 and the characteristics of the non-responders at each wave, it became clear that in order to have a good number of very low attainers in the sample at wave 4, we would need to include a large number at the first wave. This would actually represent something approaching a third of the total national number of pupils with no qualifications since, thankfully, this is a rare outcome.

Table 3: Optimum sample method required for Cohort 13 if it had been face to face

	Actual attainment distribution for 2003	Initial sample needed to ensure a representative sample at 4th wave
	percentages	percentages
None	3.5	14.7
1-4 D-G	3.4	8.1
5+ D-G	16.9	25.5
1-4 A*-C	24.1	26.0
5-7 A*-C	15.8	11.2
8+ A*-C	36.2	14.5

Source: Calculations based on Youth Cohort Study, Cohort 12, waves 1-4, 2004-2007

A further consideration in the planning for Cohort 13 of YCS was that LSYPE appeared to have been successful in avoiding the usual problems of differential non-response. The table below shows responses at wave 1, by some of the main categories across which response often varies considerably. The intention was to lift response rates to 70 per cent and to reduce the differential non-response which had been so clear in YCS. Despite the planned approach, we still expected to see significant levels of differential non-response – varying according to ethnicity and to geographical area, for example. However, the wave 1 patterns of response were remarkably consistent. It is unusual for surveys in the UK to have such consistent response rates; often response is poor amongst certain ethnic groups and in London, for example.

Table 4: Response rates for wave 1 of LSYPE, by characteristics

	full response	partial response	Non response
White	67	9	25
Black	56	12	33
Asian	62	15	23
Other	58	10	32
Male	65	10	25
Female	64	10	26
FSM	59	13	28
Non FSM	66	9	25
Independent	56	11	32
Maintained	64	10	26
London	60	9	31
Other England	65	10	25
Overall	64	10	26

Note: partial response usually means we could not interview both parents
Source: LSYPE, wave 1, 2004

Putting these two pieces of information together (the difficulties of ensuring a representative sample from our existing method and the potential benefits of moving to a face to face methodology), it was a natural consequence to move YCS to a face to face survey for the first wave of Cohort 13. This has been successful with a much improved response rate overall and much less evidence of differential non response. The final response rate was around 70 per cent, compared to below 50 per cent for the postal methodology of the previous cohort. There was some difference according to characteristics such as attainment, but much less than previously. Although cost considerations meant that our sample had to be much smaller, the higher response and more even non-response make this a price worth paying. The issued sample under face to face interviewing was only a fraction of what it would have been with postal interviewing and although even the achieved sample was also significantly smaller, the removal of serious bias is a major benefit.

Another immediate benefit of administrative data is that sample design is a much more exact science. The issued sample can include precise numbers in any given category included in the administrative data and precise boosts can be implemented. Most such surveys in England contain boosts for young people in Minority Ethnic Groups since ethnicity is recorded as standard on the administrative datasets. For LSYPE, the initial design included boosts for the 6 main Minority Ethnic Groups (Indian, Pakistani, Bangladeshi, Black African, Black Caribbean and Mixed).

Using administrative data during surveys

While the above improvements are concerned with the early stages of planning for a study, administrative data also have their uses as a study progresses. One obvious example is that the availability of administrative data can reduce the need for questions. In both LSYPE and YCS, we no longer have to ask respondents about standard academic qualifications they have achieved. This may seem like a minor point, but in fact the questions on qualifications were some of the most complex and, in some cases, monotonous for respondents to complete. It is thought that being able to remove this section of questions from our questionnaires had a significant effect on response.

In the case of LSYPE, another way in which administrative data can be used emerged later on in the study. On analysing wave 1 data, although patterns of individual level response were generally even across groups of interest, it became clear that there was something of a problem in a different form of non-response. In order to contact young people and their parents, the study team had to contact schools to request personal details for the chosen sample members since, at that time, address details were not held on the administrative record. The level of school response in this first stage of sampling was satisfactory overall (over 74 per cent). However, school response was somewhat biased in that non-response was greater amongst schools in inner city, deprived areas; 67 per cent in deprived areas compared to 78 per cent in non deprived areas. These deprived areas tend to have larger concentrations of certain groups and, in particular, it became clear that we would not be able to meet our targets for those of Black origin (both Black African and Black Caribbean). By the time that this problem came to light, address details had been added to the schools census administrative record (one of the main reasons being to make it easier to draw samples). So, we were able to reissue those sample members from these particular ethnic groups in order to restore numbers from Wave 4 onwards.

The differences between non-response and attrition

In most, if not all, longitudinal studies the drop out between waves will be less than the initial levels of non-response. This raises the possibility that the characteristics of those dropping out at different times may be different. As an example of what we might see, the following table shows drop out at different stages of YCS Cohort 12. This presents data based on that already seen in earlier tables, but looks specifically at the percentage of each group that drops out at each wave. This highlights the differences between initial non response and attrition in the later waves. There are a couple of patterns that hold for most disaggregations. Firstly, drop out is always less at wave 2 than wave 1, and much the same at wave 3 as it is at wave 2. However, drop out at wave 4 is generally higher than at wave 3, though interestingly this is not true for the very low attaining group. Secondly, not surprisingly, high non response at wave 1 is associated with higher rates of drop out in later waves. However, some gaps close considerably – compare the difference between drop out of Asian and Black young people in waves 2 and 3 for example.

Table 5: Drop out rates at each wave of YCS cohort 12, by characteristics

	Wave 1	Wave 2	Wave 3	Wave 4
GCSE attainment				
None	81	52	47	42
1-4 D-G	79	50	41	44
5+ D-G	70	41	39	46
1-4 A*-C	62	39	34	39
5-7 A*-C	53	32	33	36
8+ A*-C	37	23	25	33
Female	48	28	29	33
Male	60	34	31	40
Non FSM	52	29	29	35
FSM	69	41	35	43
Asian	55	29	29	32
Black	69	32	32	41
Other ethnic group	58	33	28	33
Unclassified	57	31	34	41
White	53	30	30	36
Total	54	30	30	36

Source Youth Cohort Study, Cohort 12, waves 1-4, 2004-2007

It is important to consider what we would know about such cases in the absence of administrative data. Without such knowledge, the table would look like the Table 6 below. The key difference is that we know very little about non-response at wave 1. It is generally possible to collect equivalent data in subsequent waves (though, in this case, it would not be possible to collect information on whether or not the young person was eligible for Free School Meals), but information on those not responding at wave 1 is gone for good. While Table 5 offers some reassurance that, in the case of YCS, patterns of attrition are reasonably well matched to patterns of non-response, this may not always be the case. Under other circumstances, such as a change in mode between waves 1 and 2, it would be more likely that important differences would be hidden in the absence of administrative data.

Table 6: Table 5 as it would look without the use of administrative data

	Wave 1	Wave 2	Wave 3	Wave 4
GCSE attainment				
None	?	52	47	42
1-4 D-G	?	50	41	44
5+ D-G	?	41	39	46
1-4 A*-C	?	39	34	39
5-7 A*-C	?	32	33	36
8+ A*-C	?	23	25	33
Female	?	28	29	33
Male	?	34	31	40
Non FSM	?	?	?	?
FSM	?	?	?	?
Asian	?	29	29	32
Black	?	32	32	41
Other ethnic group	?	33	28	33
Unclassified	?	31	34	41
White	?	30	30	36
Total	54	30	30	36

Source Youth Cohort Study, Cohort 12, waves 1-4, 2004-2007

It is clear that the availability and use of administrative data adds greatly to our understanding of what is going on. Even if administrative data are not used to correct the dataset through weighting, it is vital in being able to check any correction carried out.

Theory into practice – weighting in LSYPE

The use of administrative data in LSYPE has always been a major part of its design and strength. Initially administrative data were used as a sampling frame and have been used since to apply weighting to the data. As with the approach in YCS, a matrix approach was used at wave 1, comparing survey totals against actual population totals. Variables used were ethnicity, region (primarily whether in London or not), gender and attainment. Some additional calculations were needed to deal with the small numbers of young people who were not in maintained schools – since the full range of administrative data are not available for those in independent schools. In these cases, the only variables that could be used for weighting were region (London or not London) and type of school, depending on whether it was single sex or mixed. The weights from the independent and maintained schools then had to be combined and reweighted to reflect the proper proportion of young people in independent schools; around 7 per cent, compared to 4 per cent of the sample.

At wave 2, there is again the question of whether the attrition between waves 1 and 2 is similar in characteristics to the non response at wave 1. Although at the time of writing, this weighting process for wave 2 is yet to be finalised, early indications are that, for certain variables, there is actually more bias in wave 1 to 2 attrition than there was in initial non response. The following table presents some of the odds ratios and shows how although there is less drop out at wave 2 than wave 1, we should not assume that it is less of a problem. In these cases below the odds ratios change markedly – there is more differential attrition than there was differential non response at wave 1.

Table 7: Comparison of drop out odds ratios in waves 1 and 2 of LSYPE

	Wave 1	Wave 2
Reached targets at 14 in Maths (yes vs. no)	1.28	1.71
Reached targets at 14 in English (yes vs. no)	1.34	1.80
Reached targets at 14 in Science (yes vs. no)	1.34	1.82
Eligible for FSM (yes vs. no)	0.86	0.63
School with high levels of FSM (yes vs. no)	0.80	0.64

note: odds ratios measured as chance of response / chance of non response
Source: provisional analysis of LSYPE responses at waves 1 and 2

It is clear that the presence of administrative data enables us to take account of these sorts of differences at different waves in order to apply corrective weighting. As shown in table 6, in the absence of administrative data, the best option is simply to assume that patterns of initial non response mirror the patterns of attrition seen later on. However, as the above practical example shows, this assumption is not always valid.

Conclusions
The availability of administrative data can enable significant improvements to many stages of the survey process:

- Initial samples can be defined more precisely
- Fieldwork can be influenced by more detailed data
- Samples can easily be refreshed if necessary
- Questionnaire length can be reduced
- Data can be better weighted at every wave

All of these improvements are potentially valuable and although such comprehensive administrative data are not always available, it is well worth considering potential sources early on in the planning process. This involves finding out what sources may be available, considering how they might be used and, perhaps most importantly of all, ensuring that the study is legally able to use such data. In the UK, data protection laws are quite strict and it is important to know what requirements have to be met.