Institute of Education

A data driven approach to understanding and handling non-response in the Next Steps cohort

CLS working paper number 2020/5

By Richard J. Silverwood, Lisa Calderwood, Joseph W Sakshaug, George B. Ploubidis

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Corresponding author

Richard Silverwood

UCL Centre for Longitudinal Studies

r.silverwood@ucl.ac.uk

This working paper was first published in April 2020 by the UCL Centre for Longitudinal Studies.

UCL Institute of Education

University College London

20 Bedford Way

London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Institution of Education (IOE), University College London. It manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

Disclaimer

This working paper has not been subject to peer review.

CLS working papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of the UCL Centre for Longitudinal Studies (CLS), the UCL Institute of Education, University College London, or the Economic and Social Research Council.

How to cite this paper

Silverwood, R.J., Calderwood, L., Sakshaug, J.W., Ploubidis, G.B. (2020) *A data driven approach to understanding and handling non-response in the Next Steps cohort,* CLS Working Paper 2020/5. London: UCL Centre for Longitudinal Studies.

Summary

Non-response is common in longitudinal surveys, reducing efficiency and introducing the potential for bias. Principled methods, such as multiple imputation, are generally required to obtain unbiased estimates in surveys subject to missingness which is not completely at random. We present a systematic data-driven approach used to identify predictors of non-response in Next Steps, a national cohort study which follows a sample of young people from age 13-14 years to age 25 years. The identified predictors of non-response were across a number of broad categories. We found that including these predictors of non-response as auxiliary variables in multiple imputation analyses allowed us to restore sample representativeness in several different settings. We propose that such variables are included in future analyses using principled methods to reduce bias due to non-response in Next Steps.

Key words: Longitudinal surveys; Missing data; Multiple imputation; Non-response; Survey attrition.

Introduction

Non-response – when people or households are sampled but data are not gathered – is common in longitudinal surveys. Missing values due to non-response mean less efficient estimates because of the reduced size of the analysis sample, but also introduce the potential for bias since respondents are often systematically different from non-respondents (1). In the present paper we focus on unit non-response (when an eligible sample member fails to respond at all or does not provide enough information for the response to be deemed usable (2)) at a given wave of data collection rather than item non-response (wherein the sample member responds but does not provide a usable response to a particular item or items (2)), though the statistical issues are the same in both cases (3). Continued unit non-response at subsequent waves of a longitudinal survey results in sample attrition. There is mounting evidence that the extent of sample attrition in longitudinal studies has increased over time (4), so appropriate handling of missing data in this setting is becoming ever more important.

Missing data are typically characterised by their corresponding missing data mechanism: (i) missing completely at random (MCAR), meaning that missingness does not depend on either observed or unobserved values (i.e. is completely at random); (ii) missing at random (MAR), meaning that, given the observed values, missingness does not depend on unobserved values; or (iii) missing not at random (MNAR), meaning that missingness depends on unobserved (and possibly observed) values (5, 6). A complete-case analysis (CCA; one restricted to study participants with complete data) is valid if data are MCAR, but also under MNAR if missingness is independent of the outcome variable given the covariates in the model (7). If data are MAR then popular analysis approaches include inverse probability weighting (8, 9) and multiple imputation (MI) (3, 6, 10), the latter of which is the focus of the present paper. In MI the analyst specifies an appropriate imputation model, from which a series of imputed datasets are created. Each imputed data set is analysed using the substantive model of interest and the results are combined using standard rules (6), resulting in standard errors that incorporate the variability in results between the imputed data sets. In this way, uncertainty about the missing data is appropriately accounted for in the inference. Over recent years, MI has been widely adopted because it is practical for applied researchers in a wide range of settings and can be undertaken using standard statistical software (3).

We focus our attention on unit non-response in Next Steps, a national cohort study which follows a sample of young people age 13-14 years to age 25 years (11, 12). As in the

majority of longitudinal surveys, it seems implausible that data in Next Steps are MCAR. Interest is therefore in whether data are, for a given analysis, MAR or MNAR. Given that this distinction is not empirically testable, that convenient implementations of MI are readily available, and that MI exhibits little bias under minor deviations from MAR (13), a pragmatic approach is to undertake a MI analysis having first maximised the plausibility of the MAR assumption. In the analysis of longitudinal studies, as in other settings, it is acknowledged that the imputation model should include both variables that are associated with the probability of data being missing and variables that are predictive of variables that are subject to missingness (14). We capitalise on the rich data available in earlier waves of Next Steps and present a systematic data-driven approach used to identify predictors of unit non-response. Inclusion of these variables (alongside others) in subsequent MI analyses has the potential to maximise the plausibility of the MAR assumption. Consequently, we also investigated whether by including these variables in a MI approach we were able to restore sample representativeness despite unit non-response.

Methods

Data

Next Steps (formerly the Longitudinal Study of Young People in England) (11, 12) is a national cohort study which follows a representative sample of young people born between 1 September 1989 and 31 August 1990. It was funded and managed by the Department for Education from inception to wave 7 (15), and is now managed by the UCL Centre for Longitudinal Studies and funded by the Economic and Social Research Council. Cohort members were recruited in February 2004 while they were in Year 9 (age 13-14 years) at English state and independent schools and pupil referral units. The sample design considered schools as the primary sampling unit and included an oversampling of deprived schools and minority ethnic groups within schools. The issued sample at baseline comprised approximately 21,000 young people with a total of 15,770 persons interviewed at baseline (wave 1). There have been eight waves of data collection, with the most recent at age 25-26 years. An additional minority supplement was added at wave 4 (age 16-17 years), though we only analyse data from the original cohort in the present study. In waves 2-7 (age 14-20 years) the issued sample consisted of cohort members who had participated at the previous wave, but at wave 8 (age 25-26) the issued sample included all cohort members who had ever participated. In the first four waves both young people and their parents were interviewed; from wave 5 only young people were interviewed. The study includes information about cohort members' education and employment, economic circumstances, family life, physical and emotional health and wellbeing, social participation and attitudes. From wave 5 onwards the cohort has utilised a sequential mixed mode (web-telephone-face-to-face) design (prior to this it was face-to-face only), though we do not consider this further here.

Exposures (predictors of non-response)

Waves 1-7 of Next Steps include a total of 1252 variables that could potentially be used as predictors of non-response at wave 8. However, many of these are so-called "routed" variables, where the question is only asked of respondents that gave a specific response to a previous question. For example, only young people who report living in an institution will be asked a subsequent question on precisely what kind of institution they live in. To avoid sample selection all "routed" variables were excluded from the analysis. We used variables derived from the young person and main parent questionnaires only to avoid selection based on the completion of the questionnaire by a second parent (usually the father). We also excluded binary variables with prevalence less than 1% and variables with greater than 50% missing data. Sets of variables relating to the same or similar concepts were examined and

reduced to a smaller number of less highly correlated variables (often a single variable containing the most information). This resulted in 868 variables that met the criteria for inclusion in the analysis. They cover all domains captured by Next Steps, including details of school and education, opinions around school and schooling, behaviour and activities outside of school, health, wellbeing and health behaviours, attitudes to work, pay and the future, indicators of individual and familial socioeconomic position, and other individual and familial demographic information. In addition to these variables we calculated a binary variable which indicated whether a cohort member had failed to respond at any one or more of waves 1-7.

Outcome (non-response)

We used a binary variable indicating non-response at wave 8. We defined non-response as participants who did not take part in the survey, either because of refusal, the survey team not being able to establish contact, or because contact was not attempted.

Variables for 'sample representativeness' analyses

To examine the performance of our proposed approach to missing data handling we considered the 'sample representativeness' of several variables under different analytic approaches. These variables were chosen as they are widely used in Next Steps research. Sample representativeness was assessed both internally, by reference to survey measures from earlier waves, and externally, using 'gold standard' reference data.

Wave 1 sociodemographic characteristics: We considered a variety of important sociodemographic characteristics observed at wave 1, relating to both the young person themselves (whether they were male, non-white British, had ever been identified as having special educational needs (SEN), or had ever been suspended from school) and to their family (whether a language other than English was the main language spoken at home, their home was rented from a Council or New Town, their father had no qualifications, their father was unemployed or looking for a job, their father was employed in routine occupation, or they were a single parent household).

Wave 1 and wave 2 household salary: We considered the gross annual household salary reported at waves 1 and 2.

University attendance by wave 8: We considered the percentage of cohort members reporting that they had ever been to university by wave 8. This was selected as an important indicator of sample representativeness as it is of substantive interest in this age group and

has often been used for research purposes in Next Steps. As an external benchmark, we used the Higher Education Initial Participation Rate (HEIPR), an estimate of the likelihood of a young person participating in higher education (HE) at or by a given age, based on current participation rates. It applies to English domiciled first time participants in HE at UK HE institutions (HEIs) and English, Welsh and Scottish further education (FE) colleges (16). We derived an estimated HEIPR of 40.5% for our particular cohort (those born in 1989-90) by summing the appropriate age-specific HEIPRs published annually by the Department for Education (age 17 HEIPR in 2006/07, age 18 HEIPR in 2007/08, …, age 25 HEIPR in 2014/15) (16). However, HEIPR is not identical in scope to Next Steps university attendance data as the former includes only UK HEIs (whereas non-UK HEI attendance would also be recorded in Next Steps) but does include English, Welsh and Scottish FE Colleges (which would probably not be recorded as university attendance by Next Steps cohort members). In each year between 2012/13 and 2014/15 (the only years with available data), FE colleges provided 3.6 percentage points of the HIEPR (17). We have therefore subtracted 3.6% from our estimated HEIPR to give an estimated HEI-only participation rate of 36.9%.

Analytic strategy

The approach was based on that recently undertaken in the National Child Development Survey (18). In order to identify the important predictors of wave 8 non-response, we employed a multi-stage analytic strategy using the identified 868 eligible wave 1-7 variables as input. For predictor variables at each of waves $t = 1, \ldots, 7$ separately, we proceeded as follows.

Preliminary stage: We cross-tabulated all predictor variables at wave t against non-response at wave 8, restricted to study members with complete data on all wave t predictor variables. We ensured that all predictor variables had cell size ≥ 5 by recoding as necessary.

Stage 1: We fitted univariable modified Poisson regression models (19) relating non-response at wave 8 to each predictor variable at wave t. We performed a (joint, as necessary) Wald test for each wave t predictor variable and retained those variables with $p < 0.05$. These were the wave t "stage 1 predictors" of non-response at wave 8.

Stage 2: We fitted a multivariable modified Poisson regression model relating non-response at wave 8 to all wave t stage 1 predictor variables. We performed a (joint, as necessary) Wald test for each wave t predictor variable and retained those variables with $p < 0.05$. These were the wave t "stage 2 predictors" of non-response at wave 8.

We repeated the above process for each of waves $t = 1, \ldots, 7$ and pooled all the wave 1-7 stage 2 predictors.

Stage 3: Incomplete records become more prevalent as more waves are considered, so in order to appropriately handle missing data when relating wave 8 non-response to wave 1-7 stage 2 predictors we utilised a MI approach. The imputation model included all wave 1-7 stage 2 predictors, wave 8 non-response and the initial survey design weights. (We performed sensitivity analyses in which the functional form of the initial survey design weights were different (quadratic, categorical) and in which the imputation models were not weighted, but the ultimate set of wave 1-7 predictors of wave 8 non-response were almost identical (results not shown). We also tried fitting the imputation model separately within strata defined by the initial survey design weights but did not achieve model convergence.) We used MI by chained equations (20-22), weighted by the initial survey design weights, and generated 50 imputed datasets. We then fitted a series of multivariable modified Poisson regression models relating non-response at wave 8 to wave 1-7 stage 2 predictor variables. To preserve the temporal sequence of the longitudinal information available in Next Steps and avoid over-adjustment from conditioning on variables on the causal pathway between a given predictor and wave 8 non-response, we fitted models in which the probability of wave 8 non-response was modelled as a function of stage 2 predictors from a given wave adjusted for all identified stage 2 predictors from previous waves only (i.e. not for any variables from subsequent waves). Thus, for example, the model for wave 1 predictors featured no adjustment, and the model for wave 5 predictors was adjusted for wave 1-4 predictors only. This approach ensures that in each model we are adjusting for all the earlier variables in Next Steps potentially associated with wave 8 non-response, since these are precisely what were identified in stage 2. We appropriately accounted for the survey structure in each of the models. For variable selection in this stage we used a more stringent criterion of $p < 0.001$, with the resultant wave 1-7 variables forming our ultimate set of predictors of wave 8 non-response.

Although our proposed variable selection approach allows us to identify a set of the strongest wave 1-7 predictors of wave 8 non-response, we acknowledge that the precise *p*-values chosen to act as cut-offs are essentially arbitrary. We therefore explored how changing the stage 3 selection criterion affected the resultant set of predictor variables in two sensitivity analyses: in the first, we used a cut-off of $p < 0.01$, and in the second we used a combination of *p*-value and estimated effect magnitude, requiring $p < 0.05$ *and* a risk ratio

(RR) > 1.1 or < (1/1.1) (for categorical variables, any single between-category RR reaching this threshold was considered sufficient).

'Sample representativeness' analyses

Once the wave 1-7 predictors of wave 8 non-response were identified, we performed a number of subsequent analyses to assess the performance of our proposed approach to handling non-response in Next Steps. We investigated whether including the identified predictors of non-response in imputation models allowed us to reliably estimate the distributions of a number of variables of interest. In each imputation model we included: (i) analysis variables of interest (i.e. wave 1 sociodemographic characteristics, wave 1/2 household salary or university attendance by wave 8), (ii) a selection of wave 1 auxiliary variables relating to socioeconomic position and demographics (listed in full in Methods S1, Supplementary Material), plus the initial survey design weights, and (iii) the identified wave 1-7 predictors of wave 8 non-response. In supplementary analyses we included only (i) and (ii) in the imputation models in order to assess the added value of including the wave 1-7 predictors of wave 8 non-response. We again used MI by chained equations weighted by the initial survey design weights and generated 50 imputed datasets. We appropriately accounted for the survey structure in each of the analyses, with MI estimates weighted by the initial survey design weight.

Wave 1 characteristics: We estimated the percentage of cohort members reporting a number of characteristics observed at wave 1. We estimated these percentages on cohort members with data on these characteristics at wave 1, firstly by using all available data on each wave 1 characteristic of interest and secondly by using data on each wave 1 characteristic of interest from wave 8 respondents only, using MI (and CCA for comparison). If the MI approach using only data on wave 8 respondents provides comparable estimates to those using all available data at wave 1 it suggests that the identified predictors of wave 8 non-response are able to restore sample representativeness despite attrition between waves 1 and 8. As some wave 1 auxiliary variables may be highly correlated with the wave 1 characteristics of interest (see Methods S1, Supplementary Material), we performed a sensitivity analysis in which such wave 1 auxiliary variables were excluded.

Wave 1 and wave 2 household salary: In a similar analysis, we estimated mean wave 1 and wave 2 gross annual household salary. We estimated these quantities on cohort members who reported wave 1 and wave 2 gross annual household salary data, firstly by using all available data on wave 1 and wave 2 household salary, and secondly by using data on wave 1 and wave 2 household salary from wave 8 respondents only, using MI (and CCA for

comparison). If the MI approach using only data on wave 8 respondents provides comparable estimates to those using all available data at wave 1 and wave 2 it again suggests that the identified predictors of wave 8 non-response are able to restore sample representativeness despite attrition between waves 1/2 and 8.

University attendance by wave 8: We estimated the percentage of cohort members reporting that they had “ever been to university” by wave 8 using MI (and CCA for comparison). The resultant percentages were compared with the externally estimated HEI-only participation rate. If the MI estimate using only data from wave 8 respondents is comparable to the calculated HEIPR this provides some external validation of our approach.

All analyses were conducted using Stata version 15 (StataCorp, 2017, College Station, TX).

Results

Predictors of wave *8 non-response*

A total of 7569 out of 15,770 (48.0%) original cohort members (i.e. excluding the additional minority sample at wave 4) participated in wave 8 of Next Steps. Following the outlined approach, we identified 21 wave 1-7 predictors of wave 8 non-response (Fig. 1). These variables are reported, along with their estimated associations with wave 8 non-response, in Table 1. The strongest predictor of wave 8 non-response overall was non-response at previous waves, with previous non-responders almost 90% more likely not to respond at wave 8 than those with complete response up to wave 7 (RR 1.87, 95% confidence interval (CI) 1.79, 1.95). Other predictors of wave 8 non-response were: the cohort member being male, their parents not knowing where they are going when they went out in the evening, not having been upset by name-calling in the last 12 months, infrequent use of a home computer to play games, not having played a musical instrument in the last four weeks, living in a rented home, being unable to access the internet from home (all wave 1), the school having contacted their parents about their behaviour, not feeling under strain recently (wave 2), smoking, having younger parents (wave 3), regularly going to nightclubs, being unwilling to have their details passed on to the Department for Work and Pensions (wave 4), having moved since the previous interview, feeling that there are specific groups of people that are usually treated better by the government than people like them, thinking that their teachers expected them to do less well in their exams than their peers, no longer being in full-time education (wave 5), not having spoken to a teacher for information, advice and guidance about the future (wave 6), and unwillingness to answer questions on sexual experiences (waves 6 and 7).

In sensitivity analyses exploring how changing the stage 3 selection criterion affected the resultant set of wave 1-7 predictor variables we found that relaxing the threshold to $p < 0.01$ identified 28 variables (the 21 identified under our primary approach, plus a further 7) and the p-value and RR-based criterion identified 30 variables (19 overlapping with those under the primary approach, plus a further 11) (Table S1, Supplementary Material).

Wave 1 characteristics

The percentages of cohort members reporting a number of characteristics observed at wave 1 estimated using different methods are reported in Table 2. The number of cohort members with available data on each of the wave 1 characteristics of interest varied between 9997 and 15,663. The number of cohort members who were respondents at wave 8 and had available data on each of the wave 1 characteristics of interest varied between 5186 and

7523. For each wave 1 characteristic of interest this was approximately 50% of the available data at wave 1. The percentage of cohort members with each wave 1 characteristic of interest calculated using data from wave 8 respondents only was underestimated relative to the percentage calculated using all available wave 1 data when using CCA (for example, 45.0% vs. 51.5% male, 12.8% vs. 14.1% non-white British). When using MI including both wave 1 auxiliary variables and wave 1-7 predictors of wave 8 non-response the percentages were close to the percentages calculated using all available wave 1 data (for example, 14.3% vs. 14.1% non-white British, 21.8% vs. 21.5% ever identified as having SEN) with the exception of being male (46.6% vs. 51.5%). In the supplementary analysis which excluded the wave 1-7 predictors of wave 8 non-response from the imputation model the percentages were still close to the percentages calculated using all available wave 1 data for some variables (for example, 14.2% vs. 14.1% non-white British), but many variables they were further away than when also including the wave 1-7 predictors of wave 8 non-response (for example, 20.1% vs. 21.5% ever identified as having SEN) (Table S2, Supplementary Material). In the sensitivity analysis in which wave 1 auxiliary variables which may have been highly correlated with the wave 1 characteristics of interest were excluded from the imputation model (see Methods S1, Supplementary Material for full details) the results for most wave 1 characteristics were very similar to those using the full set of wave 1 auxiliary variables suggesting that high levels of correlation, if present, were not unduly affecting the results (Table S2, Supplementary Material). For a small number of wave 1 characteristics, for example being non-white British, the difference was more substantial, but the results using the subset of wave 1 auxiliary variables were still closer to the results using the full set of wave 1 auxiliary variables than to those from the CCA, suggesting that the good performance of the proposed MI approach was largely driven by factors other than such correlations.

Wave 1 and wave 2 household salary

The number of cohort members who reported household salary data at waves 1 and 2 were 6,927 and 7,612, respectively. Of these, 3,653 (53%) and 4,198 (55%), respectively, were also respondents at wave 8. Mean household salary was estimated to be £33,022 (95% CI £31,927, £34,118) at wave 1 and £35,676 (95% CI £34,740, £36,613) at wave 2 using all available data (Figure 2 and Table S3, Supplementary Material). When restricting analysis to wave 8 respondents, CCA overestimated the observed wave 1 and wave 2 means (£34,756 and £37,560, respectively). However, the MI estimates (£32,673 and £36,875, respectively) were more consistent with the observed means, particularly for wave 1 household salary. In the supplementary analysis which excluded the wave 1-7 predictors of wave 8 non-response

from the imputation model the estimated means were still close to the estimates using all available data (Table S3, Supplementary Material).

University attendance by wave 8

Of the 15,770 cohort members, 7,569 (48%) had data on university attendance by wave 8, with 44.5% (95% CI 42.9%, 46.2%) of these reporting having attended university (CCA; see Table 3). Using MI the estimated university attendance by wave 8 was 38.2% (95% CI 36.7%, 39.7%), closer to the calculated adjusted HEIPR of 36.9%. In the supplementary analysis which excluded the wave 1-7 predictors of wave 8 non-response from the imputation model the estimated university attendance by wave 8 was 41.4% (95% CI 39.3%, 42.9%), approximately halfway between the CCA and full MI estimates (Table S4, Supplementary Material).

Discussion

Summary of findings

Using a data-driven approach we have identified 21 variables from waves 1-7 of the Next Steps cohort that are predictive of wave 8 non-response. These variables were across a number of broad categories, including personal characteristics, schooling and behaviour in school, activities and behaviour outside of school, mental health and wellbeing, socioeconomic status, and practicalities around contact and survey completion.

We found that including the identified wave 1-7 predictors of wave 8 non-response in MI analyses allowed us to restore sample representativeness in a number of different settings. Analyses in which the wave 1-7 predictors of wave 8 non-response were not included in the imputation model suggested that, whilst for some analysis variables it was important to include the predictors of non-response in order to obtain reliable estimates, for other analysis variables this was not the case. Given that the missing data mechanisms underlying the different analysis variables will inevitably differ it seems plausible that for some the inclusion of the predictors of non-response may not be necessary for the MAR assumption to hold. However, in most cases the inclusion of the predictors of non-response did improve the obtained estimates, lending support to our proposal to include these variables in analyses affected by non-response. In practice, the missing data mechanism for a given variable would not be known and, since there is little disadvantage to including predictors of non-response in the analysis, we propose adopting this as the standard analytic approach.

Existing literature

Many of our identified predictors of non-response correspond to those previously identified in the literature, including prior non-response (18), being male (18, 23, 24), socioeconomic disadvantage (18, 23), childhood behavioural problems (18, 25), changing address (4, 18, 24, 26), not remaining in education for as long (24, 27), living in rented housing (4, 18), and smoking (27).

However, to our knowledge, some of our identified predictors of non-response have not previously been identified in the literature: for example, how often young people go out and whether their parents know where they are going, what they do in their spare time (computer games, musical instruments), and whether they speak to teachers for information, advice and guidance about the future. Whilst these novel findings are of interest, it is important that they be reproduced in other settings before being considered as established predictors of non-response, as chance findings are possible.

Strengths and limitations

There are many strengths to our study. We utilised a pre-specified data-driven approach to the identification of predictors of non-response. This allowed us to identify additional predictors of non-response that reliance on existing theory may have caused us to overlook, while avoiding theoretical predictors that were not of relevance in this specific study. We capitalised on the rich data available in earlier waves of this nationally representative survey. We assessed both the internal (using earlier variables within Next Steps) and external (using population-representative data) performance of our proposed MI-based approach to dealing with bias due to selective attrition.

The study also had a number of limitations. The use of a MI approach in stage 3 of the variable selection procedure meant we had to recode some variables (particularly unordered categorical variables) due to non-convergence of the imputation model, resulting in some loss of information. We included the initial survey design weights in the imputation models but were not able to include the interactions between this variable and all other variables as recommended in the literature (28) as the resultant number of parameters in the model would have led to instability. This should not have affected our point estimates but may have led to an overestimation of the MI standard error, potentially making our conclusions slightly conservative. Future work will consider multilevel multiple imputation in this context (29).

As we utilised a multi-stage variable selection procedure, the final variance estimates (i.e. for the associations between wave 1-7 predictors and wave 8 non-response in the stage 3 multivariable model) will tend to be downwardly biased (30), potentially leading to smaller p-values and hence false-positive inclusions within our ultimate set of predictors of non-response. However, since our $p < 0.001$ criterion is to some extent arbitrary, this is not a major limitation.

As noted, the HEIPR is not identical in scope to Next Steps university attendance data. We made an ad-hoc adjustment to address the inclusion on FE college attendance in the HEIPR, which may have introduced some error, but the exclusion of non-UK HEI attendance in the HEIPR remained unaddressed. However, this would be expected to contribute only a very small proportion of all university attendance, so any underestimation is unlikely to be substantial.

A further complexity that we have not addressed is the sequential mixed mode (web-telephone-face-to-face) design utilised since wave 5 of Next Steps. Mode effects may

plausibly have affected the values of the wave 5-7 variables (31, 32), but given the strength of association required in the identification of predictors on non-response, such differences are unlikely to have unduly affected our findings, meaning that this is not a major limitation. A related consideration is whether the values of wave 8 variables among the non-respondents should be imputed as if observed under a specific hypothetical mode.

Future work

The analysis regarding university attendance could be improved by the inclusion of auxiliary variables relating to academic attainment. This will be undertaken using data from the linked National Pupil Database.

The present study focussed on unit non-response at wave 8 of Next Steps, but for analyses using only data from earlier waves it would be instructive to identify predictors of non-response at these waves. Whilst the predictors of non-response at earlier waves are likely to be similar to those at wave 8, we plan to repeat the process outlined in this paper for all earlier waves for completeness. Similarly, the process will need to be repeated as further waves of Next Steps data are collected. We also plan to apply a similar procedure within the 1970 British Cohort Study (33) and the Millennium Cohort Study (34).

Conclusions

We have described and demonstrated the use of a data-driven approach to identify predictor variables of non-response in a longitudinal cohort study. Inclusion of these variables in subsequent analyses allowed us to overcome the bias due to selective attrition of the cohort sample. Our identification of these variables will allow users of the cohort to similarly deal with bias due to selective attrition in their analyses, using MI or other principled methods. More broadly, our data-driven approach to this issue could be used as a model for investigations in other longitudinal studies.

References

1. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ: John Wiley & Sons, Inc.; 2004.
2. Cohen MP. Unit Nonresponse. In: Lavrakas PJ, editor. Encyclopedia of Survey Research Methods. Thousand Oaks, California: Sage Publications, Inc; 2008.
3. Carpenter JR, Kenward MG. Multiple Imputation and its Application. Chichester, UK: John Wiley & Sons, Ltd; 2013.
4. Watson N, Wooden M. Identifying factors affecting longitudinal survey response. In: Lynn P, editor. Methodology of Longitudinal Surveys. Chichester: Wiley; 2009. p. 157-82.
5. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
6. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Third Edition. Hoboken, NJ: Wiley; 2020.
7. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010;29(28):2920-31.
8. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics. 2007;141(2):1281-301.
9. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278-95.
10. Little RJA, Rubin DB. The analysis of social-science data with missing values. Sociological Methods & Research. 1989;18(2-3):292-326.
11. Calderwood L, Sanchez C. Next Steps (formerly known as the Longitudinal Study of Young People in England). Open Health Data. 2016;4(1):e2.
12. University of London. UCL Institute of Education. Centre for Longitudinal Studies. Next Steps: Sweeps 1-8, 2004-2016. [data collection]. 14th Edition. UK Data Service. SN: 5545. 2018.
13. Schafer JL, Olsen MK. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research. 1998;33(4):545-71.
14. Spratt M, Carpenter J, Sterne JA, Carlin JB, Heron J, Henderson J, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. 2010;172(4):478-87.
15. Department for Education. Youth Cohort Study & Longitudinal Study of Young People in England: The Activities and Experiences of 19 year olds: England 20102011.
16. Department for Education. Statistics: participation rates in higher education 2018 [Available from: https://www.gov.uk/government/collections/statistics-on-higher-education-initial-participation-rates.
17. Department for Education. Tables: participation rates in higher education 2006 to 2017 2018 [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/743910/Tables_participation_rates_in_higher_education_2006_to_2017.xls.
18. Mostafa T, Narayanan M, Pongiglione B, Dodgeon B, Goodman A, Ploubidis GB. Understanding non-response in the 1958 British birth cohort: A data driven approach. Submitted.
19. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6.
20. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40-9.
21. Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):576-84.
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99.

23. Hawkes D, Plewis I. Modelling non-response in the National Child Development Study. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006;169(3):479-91.
24. Behr A, Bellgardt E, Rendtel U. Extent and Determinants of Panel Attrition in the European Community Household Panel. European Sociological Review. 2005;21(5):489–512.
25. Atherton K, Fuller E, Shepherd P, Strachan DP, Power C. Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort. J Epidemiol Community Health. 2008;62(3):216-23.
26. Plewis I, Ketende SC, Joshi H. The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study. Journal of Official Statistics. 2008;24(3):365-85.
27. Young AF, Powers JR, Bell SL. Attrition in longitudinal studies: who do you lose? Australian and New Zealand Journal of Public Health. 2006;30(4):353-61.
28. Seaman SR, White IR, Copas AJ, Li L. Combining Multiple Imputation and Inverse-Probability Weighting. Biometrics. 2012;68(1):129-37.
29. Quartagno M, Carpenter JR, Goldstein H. Multiple Imputation with Survey Weights: A Multilevel Approach. Journal of Survey Statistics and Methodology. 2019.
30. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol. 2008;167(5):523-9; discussion 30-1.
31. De Leeuw ED. To Mix or Not to Mix Data Collection Modes in Surveys. Journal of Official Statistics. 2005;21(2):233–55.
32. Sakshaug JW, Cernat A, Silverwood RJ, Calderwood L, Ploubidis GB, Goodman A. Measurement Equivalence in Sequential Mixed-Mode Surveys: Evidence from the Next Steps Cohort Study. Submitted.
33. Elliott J, Shepherd P. Cohort Profile: 1970 British Birth Cohort (BCS70). International Journal of Epidemiology. 2006;35(4):836-43.
34. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). International Journal of Epidemiology. 2014;43(6):1719-25.

Tables

Table 1. Estimated risk ratios and 95% confidence intervals for predictors of non-response at wave 8 (n = 15,770).

Wave	Variable	RR	95% CI
1	Sex of the young person		
	Female	1.00	(reference)
	Male	1.30	1.26, 1.36
	How often the young person's parents know where they going when they go out in the evening		
	Always	1.06	0.96, 1.16
	Usually	1.10	1.00, 1.21
	Sometimes-never	1.19	1.07, 1.31
	Don't go out in the evening	1.00	(reference)
	Whether the young person has been upset by name-calling, including by text or email, in the last 12 months		
	No	1.08	1.04, 1.13
	Yes	1.00	(reference)
	Days per week the young person uses a home computer to play games		
	None	1.10	1.04, 1.15
	1 - 2 days	1.06	1.01, 1.11
	3 - 4 days	0.99	0.94, 1.04
	Most days (5 or more)	1.00	(reference)
	Whether the young person has played a musical instrument in the last 4 weeks		
	No	1.17	1.11, 1.22
	Yes	1.00	(reference)
	Housing tenure		
	Owned outright	1.00	(reference)
	Being bought on a mortgage/ bank loan	1.00	0.95, 1.06
	Rented/other	1.16	1.09, 1.24
	Whether the young person can access the internet from home		
	No	1.15	1.10, 1.20
	Yes	1.00	(reference)
2	Whether the young person's school have ever contacted their parents about their behaviour		
	No	1.00	(reference)
	Yes	1.13	1.09, 1.18
	How much constantly under strain the young person has felt recently		
	Not at all	1.14	1.06, 1.24
	No more than usual	1.07	0.99, 1.16
	Rather more than usual	1.02	0.94, 1.12
	Much more than usual	1.00	(reference)
3	Whether the young person ever smokes cigarettes		
	No	1.00	(reference)
	Yes	1.13	1.08, 1.19
	Age of the young person's main parent [per 10 years younger]	1.06	1.02, 1.10
4	How often the young person goes to nightclubs		
	Once a week or more	1.21	1.13, 1.29
	Less than once a week	1.12	1.06, 1.20

	Hardly ever	1.10	1.05, 1.15
	Never	1.00	(reference)
	Whether the young person gives their permission to pass on their details to the Department for Work and Pensions		
	No	1.21	1.14, 1.29
	Yes	1.00	(reference)
5	Whether the young person still lives at the same address as the previous interview		
	Yes	1.00	(reference)
	No	1.18	1.11, 1.26
	Whether there are specific groups of people that the young person feels are usually treated better by the government than people like them		
	No	1.17	1.12, 1.23
	Yes	1.00	(reference)
	How well the young person thought their teachers in Year 11 and earlier expected them to do in their exams		
	Better than most pupils in their year group	1.00	(reference)
	As well as most pupils in their year group	1.10	1.04, 1.15
	Less well than most pupils in their year group	1.13	1.05, 1.20
	Current main activity of the young person		
	Full-time Education	1.00	(reference)
	Working or part working and part college	1.17	1.11, 1.24
	Other	1.13	1.07, 1.20
6	Whether the young person has spoken to a teacher for information, advice and guidance about the future		
	No	1.11	1.05, 1.17
	Yes	1.00	(reference)
	Whether the young person is willing to answer questions on sexual experiences		
	No	1.19	1.11, 1.27
	Yes	1.00	(reference)
7	Whether the young person is willing to answer questions on sexual experiences		
	No	1.14	1.06, 1.22
	Yes	1.00	(reference)
	Previous non-response (waves 1-7)		
	Complete response	1.00	(reference)
	One or more instances of non-response	1.87	1.79, 1.95

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given wave are adjusted for previously identified potential predictors of non-response at that wave and previous waves (i.e. not at subsequent waves).

All analyses appropriately account for the structure of the survey.

Table 2. Distributions of selected wave 1 characteristics among wave 1 and wave 8 respondents.

	Wave 1 respondents			Wave 8 respondents				
					CCA		MI	
	n/N	%	95% CI	n/N	%	95% CI	%	95% CI
Young person								
Male	7852/15,431	51.5	50.2, 52.8	3321/7474	45.0	43.4, 46.7	46.6	45.0, 48.1
Non-White British	5309/15,412	14.1	13.1, 15.0	2373/7465	12.8	11.7, 13.9	14.3	13.3, 15.3
Ever identified as having SEN	2934/15,452	21.5	20.4, 22.7	1284/7461	19.4	18.2, 20.6	21.8	20.5, 23.0
Ever suspended from school	1582/14,079	11.1	10.3, 12.0	509/6871	7.3	6.6, 8.2	10.5	9.4, 11.5
Family characteristics								
Language other than English is main language spoken at home	2010/15,663	4.7	4.2, 5.2	865/7523	4.1	3.6, 4.7	4.7	4.1, 5.2
Home rented from a Council or New Town	2489/15,582	13.9	13.0, 14.9	946/7486	10.9	10.0, 12.0	14.3	13.2, 15.5
Father has no qualifications	2635/9997	19.9	18.9, 21.0	1211/5186	16.8	15.6, 18.1	18.8	17.4, 20.2
Father unemployed/looking for a job	545/11,603	3.0	2.7, 3.4	229/5934	2.3	1.9, 2.7	2.9	2.4, 3.4
Father employed in routine occupation	1254/10,166	11.5	10.7, 12.3	619/5290	10.6	9.6, 11.7	11.3	10.2, 12.3
Single parent household	3950/15,632	23.5	22.6, 24.4	1546/7519	19.5	18.5, 20.5	23.3	22.2, 24.5

CCA: complete-case analysis; CI: confidence interval; MI: multiple imputation; SEN: special educational needs.

All analyses appropriately account for the structure of the survey.

Table 3. Percentage of respondents reporting university attendance by wave 8 (n = 7,569).

	%	95% CI
CCA	44.5	42.9, 46.2
MI	38.2	36.7, 39.7

CCA: complete-case analysis; CI: confidence interval; MI: multiple imputation.

All analyses appropriately account for the structure of the survey.

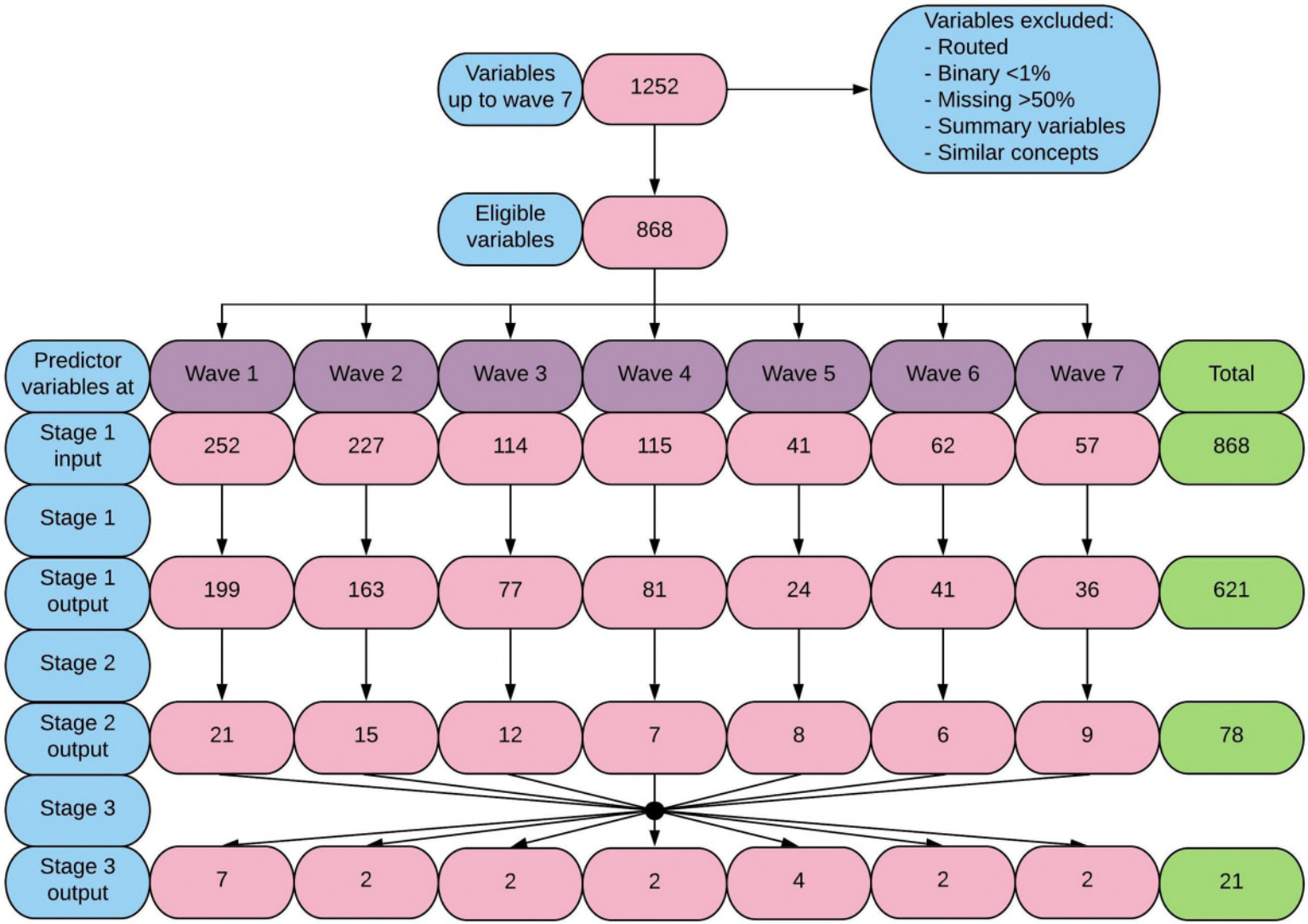

Fig. 1. Results of systematic data drive approach to non-response in Next Steps.

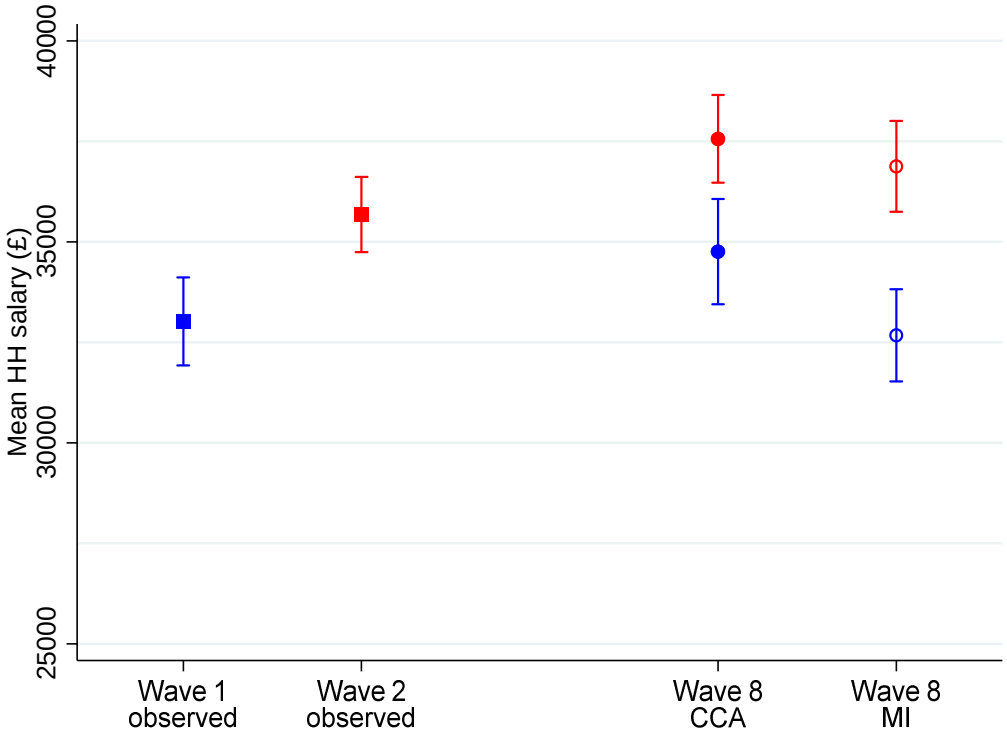

Fig. 2. Mean wave 1 (blue) and wave 2 (red) household salary estimated on cohort members who reported wave 1 and wave 2 household salary data (6,927 and 7,612, respectively) using (i) all available data and (ii) data from respondents at wave 8 only (3,653 and 4,198, respectively), using complete-case analysis (CCA) and multiple imputation (MI). All analyses appropriately account for the structure of the survey.

Supplementary Material

Methods S1

As detailed in the manuscript, we investigated whether including the identified predictors of non-response in imputation models allowed us to reliably estimate the distributions of a number of variables of interest. In each imputation model we included: (i) analysis variables of interest, (ii) a selection of wave 1 variables relating to socioeconomic position and demographics, and (iii) the identified wave 1-7 predictors of wave 8 non-response. The variables included in (ii) are listed below.

In the analyses applying the proposed missing data strategy to characteristics observed at wave 1, whether the cohort member's home was rented from a Council or New Town was derived from their housing tenure, and therefore the latter was excluded from the imputation model as it was perfectly collinear with the former. Other wave 1 characteristics were derived from variables which may be highly correlated with wave 1 auxiliary variables (for example, whether the father had no qualifications was derived from the father's highest recorded qualification, which may be highly correlated with the main parent's highest recorded qualification [though note that the father was the main parent in fewer than 20% of households]). We therefore performed a sensitivity analysis in which we excluded any such variables from the set of auxiliary variables. All such variables are marked with an asterisk below.

Variable
Sex of the young person
*Housing tenure
Whether anyone in the household has the use of motor vehicle
Whether there is a telephone in the household
Whether there is a home computer in the household
Whether the main parent or their partner currently receive child benefit
Whether the main parent or their partner currently receive invalid care allowance
Whether the main parent or their partner currently receive disability living allowance
*Whether the main parent or their partner currently receive job seekers allowance
Whether the main parent or their partner currently receive income support
Whether the main parent or their partner currently receive incapacity benefit
Whether the main parent or their partner currently receive working tax credit

Whether the main parent or their partner currently receive child tax credit

Age the main parent first left school

Whether the main parent went back into full-time education after leaving school

Whether the main parent's father ever went to university and got a degree

Whether the main parent 's mother ever went to university and got a degree

*Highest qualification held by the main parent

Age of the main parent

*Marital status of the main parent

*Employment status of the main parent

The number of dependent children in the household

*The main parent’s National Statistics Socio-economic Classification class

Whether disability limits the main parent's activity

*The main parent's ethnic group

Table S1. Identified wave 1-7 predictors of wave 8 non-response using primary and alternative approaches.

Wave	Variable	Primary approach	Alternative approach 1	Alternative approach 2
1	Sex of the young person	X	X	X
	How often the young person's parents know where they going when they go out in the evening	X	X	X
	Whether the young person has been upset by name-calling, including by text or email, in the last 12 months	X	X	
	Days per week the young person uses a home computer to play games	X	X	X
	Whether the young person has played a musical instrument in the last 4 weeks	X	X	X
	Housing tenure	X	X	X
	Whether the young person can access the internet from home	X	X	X
	Whether the main parent or their partner get involved in Parents and Teachers Associations		X	X
	Marital status of main parent		X	
	How satisfied the young person is with the sports facilities at school			X
	Whether the young person finds the work they do in lessons to be interesting to them			X
2	Whether the young person's school have ever contacted their parents about their behaviour	X	X	X
	How much constantly under strain the young person has felt recently	X	X	X
	Whether the young person has been made to hand over money or possessions in the last 12 months			X
3	Whether the young person ever smokes cigarettes	X	X	X
	Age of the young person's main parent	X	X	
	Whether the young person has been upset by name-calling, including by text or email, in the last 12 months since the last interview		X	X
	The main parent and their partner have other sources of income, e.g. rent			X
	The main parent's current National Statistics Socio-economic Classification class			X

4	How often the young person goes to nightclubs	X	X	X
	Whether the young person gives their permission to pass on their details to the Department for Work and Pensions	X	X	X
	Housing tenure			X
5	Whether the young person still lives at the same address as the previous interview	X	X	X
	Whether there are specific groups of people that the young person feels are usually treated better by the government than people like them	X	X	X
	How well the young person thought they were expected to do in their exams by their teachers in Year 11 and earlier	X	X	X
	Current main activity of the young person	X	X	X
	Whether the young person agrees that people being attacked or harassed because of their race, ethnic origin or religion is a big problem in their local area		X	X
	The young person's likelihood of voting in the next general election		X	
	Whether the young person agrees that young people today are often stopped by the police for no good reason			X
6	Whether the young person has spoken to a teacher for information, advice and guidance about the future	X	X	X
	Whether the young person is willing to answer questions on sexual experiences	X	X	X
	Whether the young person has spoken to anyone else (i.e. not friends, relatives, teachers, …) for information, advice and guidance about the future			X
7	Whether the young person is willing to answer questions on sexual experiences	X	X	X
	Previous non-response (waves 1-7)	X	X	X
	Whether the young person currently holds a driving licence		X	
	Whether the young person has ever tried cannabis		X	
	Total	**21**	**28**	**30**

Variable selection approaches differed from each other only in the criterion applied at stage 3. Primary approach: $p < 0.001$; Alternative approach 1: $p < 0.01$; Alternative approach 2: $p < 0.05$ and (risk ratio > 1.1 or risk ratio < (1/1.1)).

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given wave are adjusted for previously identified potential predictors of non-response at that wave and previous waves (i.e. not at subsequent waves).

Table S2. Distributions of selected wave 1 characteristics among wave 1 and wave 8 respondents.

	Wave 1 respondents			Wave 8 respondents								
					CCA		MI†		MI‡		MI	
	n/N	%	95% CI	n/N	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Young person												
Male	7852/15,431	51.5	50.2, 52.8	3321/7474	45.0	43.4, 46.7	44.8	43.4, 46.2	46.5	44.9, 48.1	46.6	45.0, 48.1
Non-White British	5309/15,412	14.1	13.1, 15.0	2373/7465	12.8	11.7, 13.9	14.2	13.2, 15.2	13.7	12.8, 14.7	14.3	13.3, 15.3
Ever identified as having SEN	2934/15,452	21.5	20.4, 22.7	1284/7461	19.4	18.2, 20.6	20.1	18.9, 21.2	21.7	20.5, 23.0	21.8	20.5, 23.0
Ever suspended from school	1582/14,079	11.1	10.3, 12.0	509/6871	7.3	6.6, 8.2	8.3	7.5, 9.1	10.5	9.5, 11.5	10.5	9.4, 11.5
Family characteristics												
Language other than English is main language spoken at home	2010/15,663	4.7	4.2, 5.2	865/7523	4.1	3.6, 4.7	4.7	4.2, 5.2	4.6	4.1, 5.2	4.7	4.1, 5.2
Home rented from a Council or New Town	2489/15,582	13.9	13.0, 14.9	946/7486	10.9	10.0, 12.0	13.5	12.4, 14.6	14.2	13.0, 15.4	14.3	13.2, 15.5
Father has no qualifications	2635/9997	19.9	18.9, 21.0	1211/5186	16.8	15.6, 18.1	18.3	17.0, 19.6	18.2	16.8, 19.7	18.8	17.4, 20.2
Father unemployed/looking for a job	545/11,603	3.0	2.7, 3.4	229/5934	2.3	1.9, 2.7	2.9	2.4, 3.3	2.9	2.3, 3.4	2.9	2.4, 3.4
Father employed in routine occupation	1254/10,166	11.5	10.7, 12.3	619/5290	10.6	9.6, 11.7	11.4	10.3, 12.5	11.4	10.1, 12.7	11.3	10.2, 12.3
Single parent household	3950/15,632	23.5	22.6, 24.4	1546/7519	19.5	18.5, 20.5	22.3	21.2, 23.3	23.3	22.2, 24.5	23.3	22.2, 24.5

CCA: complete-case analysis; CI: confidence interval; MI: multiple imputation; SEN: special educational needs.

MI† excludes the wave 1-7 predictors of wave 8 non-response from the imputation model.

MI‡ excludes wave 1 auxiliary variables which may be highly correlated with wave 1 characteristics from the imputation model (see Methods S1, Supplementary Material).

All analyses appropriately account for the structure of the survey.

Table S3. Mean wave 1 and wave 2 household salary (£) estimated on cohort members who reported wave 1 and wave 2 household salary data

Household salary observed at	Wave 1/2 respondents			Wave 8 respondents						
					CCA		MI†		MI	
	N	Mean	95% CI	N	Mean	95% CI	Mean	95% CI	Mean	95% CI
Wave 1	6927	33,022	31,927, 34,118	3653	34,756	33,444, 36,069	33,007	31,799, 34,215	32,673	31,523, 33,822
Wave 2	7612	35,676	34,740, 36,613	4198	37,560	36,468, 38,652	36,891	35,778, 38,004	36,875	35,744, 38,007

CCA: complete-case analysis; CI: confidence interval; MI: multiple imputation.

MI† excludes the wave 1-7 predictors of wave 8 non-response from the imputation model.

All analyses appropriately account for the structure of the survey.

Table S4. Percentage of respondents reporting university attendance by wave 8 (n = 7,569).

	%	95% CI
CCA	44.5	42.9, 46.2
MI†	41.4	39.9, 42.9
MI	38.2	36.7, 39.7

CCA: complete-case analysis; CI: confidence interval; MI: multiple imputation.

MI† excludes the wave 1-7 predictors of wave 8 non-response from the imputation model.

All analyses appropriately account for the structure of the survey.