Longitudinal and Life Course Studies • vol XX • no XX • 1–24 • © Authors 2024
Online ISSN 1757-9597 • https://doi.org/10.1332/17579597Y2024D000000010
Accepted for publication 24 January 2024 • First published online 15 February 2024

RESEARCH ARTICLE

A data-driven approach to understanding non-response and restoring sample representativeness in the UK Next Steps cohort

Richard J. Silverwood, R.Silverwood@ucl.ac.uk
Lisa Calderwood, L.Calderwood@ucl.ac.uk
Morag Henderson, Morag.Henderson@ucl.ac.uk
University College London, UK

Joseph W. Sakshaug, joe.sakshaug@iab.de
University of Warwick, UK
Institute for Employment Research, Germany
Ludwig Maximilian University of Munich, Germany

George B. Ploubidis, G.Ploubidis@ucl.ac.uk
University College London, UK

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. The inclusion of predictors of non-response in such methods, for example as auxiliary variables in multiple imputation, can help improve the plausibility of the missing at random assumption underlying these methods and hence reduce bias. We present a systematic data-driven approach used to identify predictors of non-response at Wave 8 (age 25–26) of Next Steps, a UK national cohort study that follows a sample of 15,770 young people from age 13–14 years. The identified predictors of non-response were across a number of broad categories, including personal characteristics, schooling and behaviour in school, activities and behaviour outside of school, mental health and well-being, socio-economic status, and practicalities around contact and survey completion. 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, though we acknowledge that this is unlikely to universally be the case. We propose that these variables are considered for inclusion in future analyses using principled methods to explore and attempt to reduce bias due to non-response in Next Steps. Our data-driven approach to this issue could also be used as a model for investigations in other longitudinal studies.

Keywords cohort studies • missing data • multiple imputation • non-response • sample representativeness

Key messages

- We used a data-driven approach to identify predictors of non-response in the Next Steps cohort.
- By including these variables in analyses, we could restore sample representativeness in different settings.
- These variables should be considered for inclusion in future analyses to explore and attempt to reduce non-response bias.
- Our approach could be a model for investigations in other longitudinal studies.

To cite this article: Silverwood, R.J., Calderwood, L., Henderson, M., Sakshaug, J.W. and Ploubidis, G.B. (2024) A data-driven approach to understanding non-response and restoring sample representativeness in the UK Next Steps cohort, *Longitudinal and Life Course Studies*, XX(XX): 1–24, DOI: 10.1332/17579597Y2024D000000010

Introduction

Non-response 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 (Rubin, 2004). In the present paper we focus on non-response at a given wave of data collection ('wave non-response') rather than item non-response, though the statistical issues are similar in each case (Carpenter and Kenward, 2013). Continued wave 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 (Watson and Wooden, 2009), 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: (1) missing completely at random (MCAR), meaning that missingness does not depend on either observed or unobserved values (that is, completely at random); (2) missing at random (MAR), meaning that, given the observed values, missingness does not depend on unobserved values; or (3) missing not at random (MNAR), meaning that missingness depends on unobserved (and possibly observed) values (Rubin, 1976; Little and Rubin, 2020). A complete-case analysis (CCA; one restricted to study participants with complete data on all analysis variables) is valid if data are MCAR, but also under MNAR if missingness is independent of the outcome variable given the covariates in the model (White and Carlin, 2010). If data are MAR then popular analysis approaches include inverse probability weighting (Wooldridge, 2007; Seaman and White, 2013), full information maximum likelihood (Enders, 2001) and multiple imputation (MI) (Little and Rubin, 1989; 2020; Carpenter and Kenward, 2013), the latter of which is used in the present paper. In MI the analyst specifies an appropriate imputation model, from which a series of imputed data sets are created. Each imputed data set is analysed using the substantive (analysis) model of interest and the results are combined using standard rules (Little and Rubin, 2020), 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 (Carpenter and Kenward, 2013).

We focus our attention on wave non-response at the most recent wave (Wave 8) of Next Steps, a UK national cohort study that follows a sample of young people age 13–14 years at recruitment (Calderwood and Sanchez, 2016; University of London et al, 2018). Next Steps is widely used in research relating to educational transitions and the lives of young adults. However, there has been recent interest in the representativeness of the respondents in Next Steps and how this may affect analyses of the data (Siddiqui et al, 2019), emphasising the importance of appropriate handling of non-response.

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 (Schafer and Olsen, 1998), a pragmatic approach is to undertake a MI analysis having first maximised the plausibility of the MAR assumption. This can be achieved through the inclusion of appropriate auxiliary (not in the analysis model of interest) variables in the imputation model. In the analysis of longitudinal studies, as in other settings, it is acknowledged that the imputation model should include variables that are predictive of the underlying values of variables that are subject to missingness, especially those that are also associated with the probability of data being missing (Spratt et al, 2010). 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 wave non-response. The identified set of predictors of non-response represents a pool of such variables from which researchers may draw variables that are also associated with their missingness-affected substantive variables of interest on an analysis-specific basis. Inclusion of these variables (alongside others) in subsequent MI analyses has the potential to maximise the plausibility of the MAR assumption. Consequently, we investigated whether by including these variables in a MI approach we were able to restore sample representativeness despite wave non-response. We also provide an illustrative regression example where missingness is handled using MI. Our proposed approach provides one potential solution to handling missing data, but the sensitivity of analysis findings to the specific method used for missing data handling should always be explored.

Predictors of non-response in longitudinal surveys

The primary objective of the present study was to identify predictors of non-response in the Next Steps cohort. While there is an existing literature on predictors of non-response in longitudinal surveys which we could have used as the basis for a theory-led approach to the identification of predictors of non-response in Next Steps, we chose to adopt a purely data-driven approach. This strategy allows us to potentially identify additional predictors that do not fit within the existing theory and avoids theoretical predictors that are not of relevance in the case of Next Steps. A brief review of the existing literature focusing on the aspects most relevant to the Next Steps context is provided in Literature S1, Supplementary Material. Previously

identified predictors of non-response include socio-demographic factors such as gender, ethnicity and age, marriage and cohabitation, home type and home ownership, socio-economic disadvantage, geography, residential mobility, and educational and health-related factors.

Methods

Overview of our approach

In order to meet our objective of identifying predictors of non-response at Wave 8 of the Next Steps cohort we applied a multistage, data-driven approach. The input into this process was all available variables collected between Waves 1 and 7, subject to some exclusion criteria. Application of our data-driven approach allowed us to identify a set of Wave 1–7 variables that were strongly predictive of Wave 8 non-response. We then performed a number of subsequent analyses to assess the performance of our proposed approach to handling non-response in Next Steps, considering the 'sample representativeness' of several variables under different analytic approaches. Finally, we also conducted an illustrative regression analysis.

Data

Next Steps (formerly the Longitudinal Study of Young People in England) (Calderwood and Sanchez, 2016; University of London et al, 2018) is a national cohort study that 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 (DfE, 2011), and is now managed by the UCL Centre for Longitudinal Studies. 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 ethnic 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 well-being, social participation and attitudes. From Wave 5 onwards the cohort has used 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. At Wave 8, there were 7,569 respondents out of the 15,770 persons interviewed at baseline (48.0%). Response rates for earlier waves are reported in Table S1 (Supplementary Material).

Exposures (predictors of non-response)

Waves 1–7 of Next Steps include a total of 1,252 variables that could potentially be used as predictors of non-response at Wave 8 (age 25–26). 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. 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, well-being and health behaviours, attitudes to work, pay and the future, indicators of individual and familial socio-economic 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; for the purposes of our analyses this was considered as a variable observed at Wave 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, because of refusal, the survey team not being able to establish contact, or because contact was not attempted (for example because the cohort member was known to be in prison or deceased; see Calderwood, 2018 for full details). Cohort members who had died prior to Wave 8 or were no longer living in the UK are not in the target population so would ideally have been excluded from the analysis. However, the information in relation to this, to the extent that it is reliably known at all, is not available for research purposes and we were therefore unable to make these exclusions. We would expect the numbers of study members affected by this, particularly by mortality, to be low in this young cohort and therefore this is unlikely to make a meaningful difference to the findings of the study.

In the survey literature, participation is often considered as involving two sequential events – contact and response – with predictors for each event sometimes considered separately (Watson and Wooden, 2009). We have chosen to combine these events in our definition of non-response as our aim is the identification of variables predictive of cohort members being absent from subsequent analyses due to having incomplete data, so distinguishing between the two events is unnecessary for our purposes.

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' population reference data.

We considered a variety of important socio-demographic 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 corporation (homes run by these corporations were later handed to councils), their father had no qualifications, their father was unemployed or looking for a job, their father was employed in a routine occupation, or they were a single parent household). We also considered the gross annual household salary reported at Waves 1 and 2.

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 (DfE, 2018). We derived an estimated HEIPR of 36.9% for our particular cohort (details in Methods S1, Supplementary Material).

Analytic strategy

The approach was based on that recently undertaken in the National Child Development Study (Mostafa et al, 2021). In order to identify the important predictors of Wave 8 non-response, we employed a multistage analytic strategy using the identified 868 eligible Wave 1–7 variables as inputs. For predictor variables at each of Waves $t = 1, ..., 7$ separately, we proceeded as follows.

Preliminary stage: we cross-tabulated all binary/categorical 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 to reduce sparse cells across response/non-response and independent variables.

Stage 1: we fitted a series of univariable (single independent variable) modified Poisson regression models (Zou, 2004) relating non-response at Wave 8 to each individual predictor variable at Wave t. We used modified Poisson regression due to the ease of interpretation of the risk ratio (RR) and to avoid issues related to non-collapsibility of the odds ratio. Only cohort members with available data on a given predictor variable were included in the model for that variable. We performed a (joint, as necessary) Wald test for each Wave t predictor variable and retained those variables with $p < .05$. 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. Only cohort members with available data on all Wave t stage 1 predictor variables were included in the model for Wave t. We performed a (joint, as necessary) Wald test for each Wave t predictor variable and retained those variables with $p < .05$. These were the Wave t 'stage 2 predictors' of non-response at Wave 8. We repeated this process for each of Waves $t = 1, ..., 7$.

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 used a MI approach. The imputation model included all Wave 1–7 stage 2 predictors and Wave 8 non-response. The same set of imputed data sets were used for all stage 3 analyses. In this and all subsequent implementations of MI there were a number of common features: the initial survey design weights were included in the imputation model; we used MI by chained equations (Azur et al, 2011; White et al, 2011; Harel et al, 2018); the imputation model was weighted by the initial survey design weights; we generated 50 imputed data sets; for each variable with missing data either linear, logistic, ordinal logistic or multinomial logistic regression was used as appropriate; and convergence for each imputed variable was assessed using trace plots of the mean and standard deviation. Following imputation of missing values, we fitted a series of multivariable modified Poisson regression models relating non-response at Wave 8 to Wave 1–7 stage 2 predictor variables. We were careful to ensure that we preserved the temporal sequence of the longitudinal information available in Next Steps and avoided over-adjustment from conditioning on variables on the causal pathway between a given predictor and Wave 8 non-response. We therefore fitted models in which 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 (that is, not for any variables from subsequent waves) by including these variables in the model. 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 complex sample design (strata, primary sampling units and initial survey design weights) in each of the models. For variable selection in this stage we used a more stringent criterion of $p < .001$, 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 $p < .01$, and in the second we used a combination of p-value and estimated effect magnitude, requiring $p < .05$ *and* a 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 several variables of interest. The analysis variables of interest (Wave 1 socio-demographic characteristics, Wave 1/2 household salary or university attendance by Wave 8) were analysed using three separate MI implementations. In each imputation model we included: (1) the analysis variable(s)

of interest, (2) a selection of Wave 1 auxiliary variables relating to socio-economic position and demographics (listed in full in Methods S2, Supplementary Material), and (3) the identified Wave 1–7 predictors of Wave 8 non-response. In supplementary analyses we included only (1) and (2) in the imputation models in order to assess the added value of including the Wave 1–7 predictors of Wave 8 non-response. We appropriately accounted for the complex sample design in each of the analyses, with MI estimates weighted by the initial survey design weight.

Prior to conducting the analyses regarding restoring sample representativeness, we first explored the associations between the variables in (1) and (2) and the variables in (1) and (3) to assess whether the variables in (2) and (3) were sufficiently well associated with the analysis variables in (1) to constitute potentially useful auxiliary variables. We appropriately accounted for the complex sample design in each of the analyses, with MI estimates weighted by the initial survey design weight.

We estimated the percentage of cohort members reporting a number of socio-demographic characteristics observed at Wave 1. We estimated these percentages using three approaches: (1) using all available data on each Wave 1 socio-demographic characteristic; (2) using Wave 1 socio-demographic characteristic data from Wave 8 respondents only in a CCA; (3) using Wave 1 socio-demographic characteristic data from Wave 8 respondents only (that is, recoding the values of the Wave 1 socio-demographic characteristics to be missing in Wave 8 non-respondents) but using MI to handle the 'missing' data among Wave 8 non-respondents. Approach (1) therefore forms the known 'truth' within the sample. Comparison of (2) with (1) allows us to examine the extent of bias due to non-response at Wave 8. Comparison of (3) with (1) and (2) allows us to examine the extent to which our proposed approach for handling non-response overcomes the identified bias. If (3) and (1) are comparable (and both differ from (2)) it suggests that the MI approach using the identified predictors of Wave 8 non-response and Wave 1 auxiliary variables was 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 socio-demographic characteristics of interest (see Methods S2, Supplementary Material), we performed a sensitivity analysis in which such Wave 1 auxiliary variables were excluded. We emphasise that we are only imputing Wave 1 socio-demographic characteristic data among Wave 8 non-respondents who were in actuality observed at Wave 1 for demonstration purposes – in real applications one would always use the observed values of these variables.

We estimated mean Wave 1 and Wave 2 gross annual household salary using the three estimation approaches already described in relation to Wave 1 socio-demographic characteristics.

We estimated the percentage of cohort members reporting that they had 'ever been to university' by Wave 8 using CCA and MI. The resultant percentages were compared with the externally estimated HEIPR. 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.

Illustrative regression analysis

We also conducted an illustrative regression analysis in which we examined the association between the highest qualification held by the cohort member's main

parent and the cohort member ever having attended university by Wave 8. Modified Poisson regression was used to estimate unadjusted and adjusted (for sex of the young person, age of the main parent and the main parent's ethnic group) models. Both CCA (requiring observed data on all variables in the adjusted model) and MI analysis (allowing inclusion of the whole Next Steps sample of 15,770) were conducted. The imputation model for the MI analysis included: (1) analysis variables of interest (that is, those included in the adjusted model), (2) the previously described Wave 1 auxiliary variables relating to socio-economic position and demographics and (3) the identified Wave 1–7 predictors of Wave 8 non-response. Prior to conducting this analysis, we considered the previously estimated associations between the cohort member ever having attended university by Wave 8 (the only variable in the analysis model with substantial missingness) and the variables in (2) and (3) to assess whether the variables in (2) and (3) constituted potentially useful auxiliary variables. We emphasise that this analysis is for illustrative purposes only and the substantive findings should not be meaningfully interpreted.

All analyses were conducted using Stata version 17 (StataCorp, 2017).

Results

Predictors of Wave 8 non-response

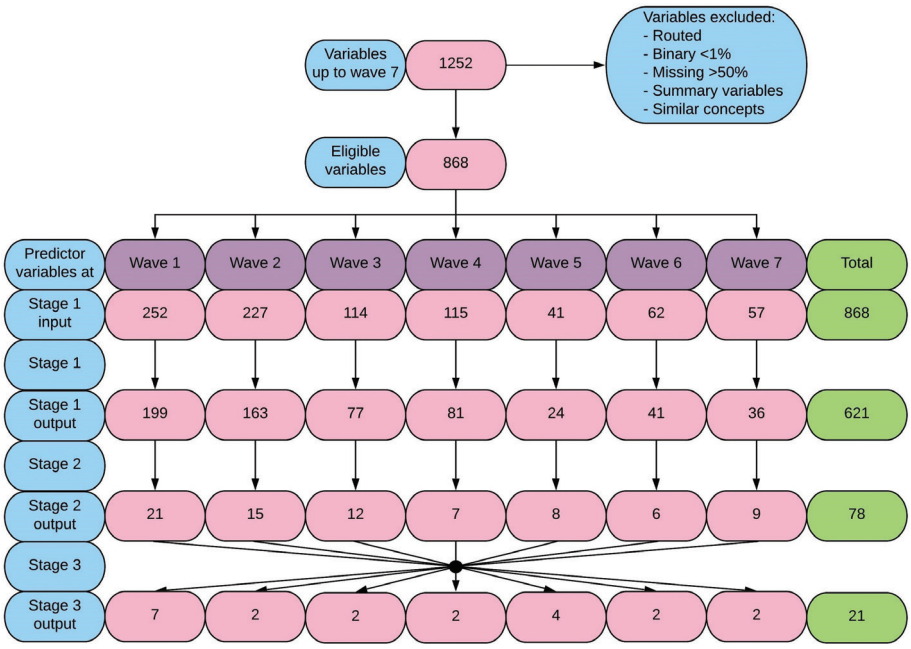
Participation in Waves 1–8 of Next Steps is shown in Table S1 (Supplementary Material). A total of 7,569 out of 15,770 (48.0%) original cohort members (that is, excluding the additional minority ethnic sample at Wave 4) participated in Wave 8. Following the outlined approach, we identified 21 Wave 1–7 predictors of Wave 8 non-response (Figure 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).

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 $p < .01$ 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 S2, Supplementary Material).

'Sample representativeness' analyses

The percentages of cohort members reporting the selected Wave 1 socio-demographic characteristics estimated using different methods are reported in Table 2. The number of cohort members with available data on each of the Wave 1 socio-demographic characteristics of interest varied between 9,997 and 15,663. The number of cohort members who were respondents at Wave 8 and had available data on each of the Wave 1 socio-demographic characteristics of interest varied between 5,186 and 7,523. 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

Figure 1: Results of systematic data-driven approach to non-response in Next Steps



relative to the percentage calculated using all available Wave 1 data when using CCA (for example, 45.0% versus 51.5% male, 12.8% versus 14.1% non-White British).

For most of the Wave 1 socio-demographic characteristics there was strong evidence of associations with virtually all of the predictors of non-response at Wave 8 (Table S3, Supplementary Material), though for some variables this was less often the case. The Wave 1 socio-demographic characteristics were similarly seen to be consistently associated with the Wave 1 auxiliary variables (Table S4, Supplementary Materials), with the one exception being whether the young person was male. As there was strong evidence of associations for all the auxiliary variables, they were all included in the imputation model. 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% versus 14.1% non-White British, 21.8% versus 21.5% ever identified as having SEN) with the exception of being male (46.6% versus 51.5%).

In the supplementary analysis that 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% versus 14.1% non-White British), but for many variables they were further away than when also including the Wave 1–7 predictors of Wave 8 non-response (for example, 20.1% versus 21.5% ever identified as having SEN; Table S5, Supplementary Material). In the sensitivity analysis in which Wave 1 auxiliary variables that may have been highly correlated with the Wave 1 socio-demographic characteristics of interest were excluded from the imputation model the results for most Wave 1 socio-demographic 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

Table 1: Estimated risk ratios (RR) and 95% confidence intervals (CI) 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)

(Continued)

Table 1: Continued

Wave	Variable	RR	95% CI
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)

(Continued)

Table 1: Continued

Wave	Variable	RR	95% CI
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

Notes:
Results from sequential analyses following multiple imputation 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 (that is, not at subsequent waves).
All analyses appropriately account for the complex sample design.

unduly affecting the results (Table S4, Supplementary Material). For a small number of Wave 1 socio-demographic 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.

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 S6, 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). There was strong evidence of associations between Wave 1 and Wave 2 household salary and virtually all the predictors of non-response at Wave 8 (Table S7, Supplementary Material) and Wave 1 auxiliary variables (Table S8, Supplementary Materials). The exceptions to this were the sex of the young person and whether the main parent or their partner currently received child benefit, though we retained these variables in the imputation model for completeness. The MI estimates (£32,673 and £36,875, respectively) were more consistent with the observed means than were the CCA estimates, particularly for Wave 1 household salary (Figure 2 and Table S6, Supplementary Material). In the supplementary analysis that 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 S6, Supplementary Material).

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; Table 3). There was strong evidence of associations between university attendance by Wave 8 and virtually all the predictors of non-response at Wave 8 (Table S9, Supplementary Material) and Wave 1 auxiliary variables (Table S10, Supplementary Materials). 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% (Table 3). In the supplementary analysis which excluded the Wave 1–7 predictors of Wave 8 non-response from the imputation model the estimated university attendance

Table 2: Distributions of selected Wave 1 socio-demographic characteristics among Wave 1 and Wave 8 respondents

	Wave 1 respondents			Wave 8 respondents					
				CCA			MI		
	n/N	%	95% CI	n/N	%	95% CI	N	%	95% CI
Young person									
Male	7,852/15,431	51.5	50.2, 52.8	3,321/7,474	45.0	43.4, 46.7	15,431	46.6	45.0, 48.1
Non-White British	5,309/15,412	14.1	13.1, 15.0	2,373/7,465	12.8	11.7, 13.9	15,412	14.3	13.3, 15.3
Ever identified as having SEN	2,934/15,452	21.5	20.4, 22.7	1,284/7,461	19.4	18.2, 20.6	15,452	21.8	20.5, 23.0
Ever suspended from school	1,582/14,079	11.1	10.3, 12.0	509/6,871	7.3	6.6, 8.2	14,079	10.5	9.4, 11.5
Family characteristics									
Language other than English is main language spoken at home	2,010/15,663	4.7	4.2, 5.2	865/7,523	4.1	3.6, 4.7	15,663	4.7	4.1, 5.2
Home rented from a council or new town	2,489/15,582	13.9	13.0, 14.9	946/7,486	10.9	10.0, 12.0	15,582	14.3	13.2, 15.5
Father has no qualifications	2,635/9,997	19.9	18.9, 21.0	1,211/5,186	16.8	15.6, 18.1	9,997	18.8	17.4, 20.2
Father unemployed/looking for a job	545/11,603	3.0	2.7, 3.4	229/5,934	2.3	1.9, 2.7	11,603	2.9	2.4, 3.4
Father employed in routine occupation	1,254/10,166	11.5	10.7, 12.3	619/5,290	10.6	9.6, 11.7	10,166	11.3	10.2, 12.3
Single parent household	3,950/15,632	23.5	22.6, 24.4	1,546/7,519	19.5	18.5, 20.5	15,632	23.3	22.2, 24.5

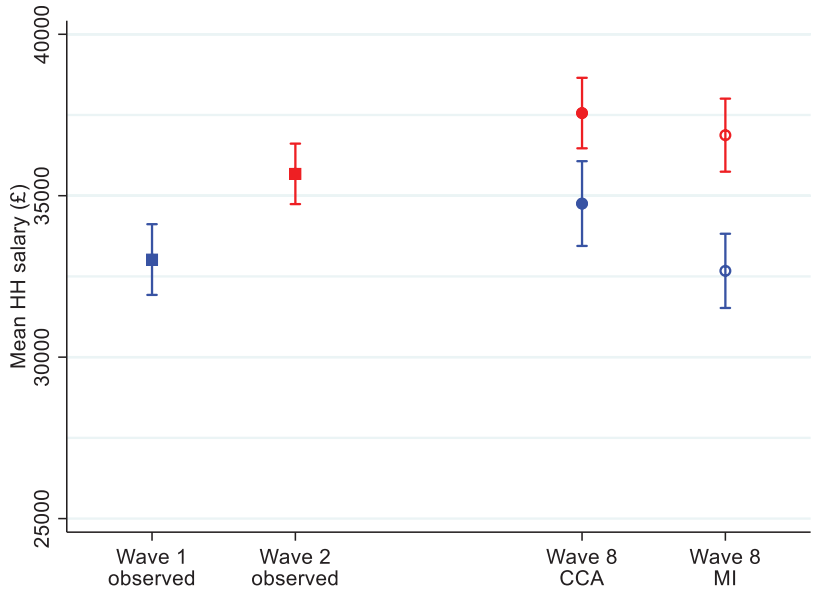
Notes:
CCA: complete-case analysis; CI: confidence interval; MI: multiple imputation; SEN: special educational needs.
n: number of cohort members with a given characteristic; N: total number of cohort members with observed data.
All analyses appropriately account for the complex sample design.

by Wave 8 was 41.4% (95% CI 39.3%, 42.9%), approximately halfway between the CCA and full MI estimates (Table S11, Supplementary Material).

Illustrative regression analysis

The vast majority of the 15,770 persons interviewed at baseline in Next Steps had observed data on the highest qualification held by their main parent (95.7%), the sex of the young person (97.9%), the age of the main parent (98.9%) and the main parent's ethnic group (98.9%). However, only 48.0% had observed data on the cohort member ever having attended university by Wave 8, meaning that the CCA sample constituted only 45.0% of the total sample. As noted previously, there was strong evidence of associations between university attendance by Wave 8 and virtually all the predictors of non-response at Wave 8 (Table S9, Supplementary Material) and Wave 1 auxiliary variables (Table S10, Supplementary Materials). In the adjusted CCA model, higher qualifications held by the cohort member's main parent were associated with higher levels of university attendance by the cohort member (Table S12, Supplementary Material). For example, having a HE qualification led to a 2.56-fold risk of university attendance relative to having no qualification (RR 2.56; 95% CI 2.30, 2.86). A similar pattern of associations was observed in the MI analysis (n = 15,770), though the magnitude of association

Figure 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 (1) all available data and (2) data from respondents at Wave 8 only (3,653 and 4,198, respectively), using complete-case analysis (CCA) and multiple imputation (MI)



Note: All analyses appropriately account for the complex sample design.

Table 3: Percentage of respondents reporting university attendance by Wave 8

	n	N	%	95% CI
CCA	3,539	7,569	44.5	42.9, 46.2
MI		15,770	38.2	36.7, 39.7

Notes:
CCA: complete-case analysis; CI: confidence interval; MI: multiple imputation.
n: number of cohort members with a given characteristic; N: total number of cohort members with observed data.
All analyses appropriately account for the complex sample design.

was generally somewhat greater, with the corresponding HE qualification RR estimated to be 2.82 (95% CI 2.53, 3.15).

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 strongly 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 well-being, socio-economic status, and practicalities around contact and survey completion.

We found that including the identified Wave 1–7 predictors of Wave 8 non-response as auxiliary variables 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, while 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. Moreover, there may be examples in which the variables subject to missingness remain MNAR even conditional on the predictors of non-response. In such cases, our proposed approach to missing data handling may not substantially reduce bias – and may even potentially exacerbate it.

We therefore suggest that in analyses of Next Steps Wave 8 data the identified predictors of non-response should be considered for inclusion as auxiliary variables. In particular, since effective auxiliary variables should be predictive of the underlying values of variables that are subject to missingness, associations with analysis variables should first be explored and variables chosen from the pool of predictors of non-response on the basis of such associations. Our identified predictors of non-response do not directly apply to other waves of Next Steps data, but may form a reasonable starting point when designing analyses of data from other waves that are subject to missingness.

For a given analysis in which the 'true' value of an estimate is unknown, it will be difficult to say to what extent application of this approach is reducing bias. However, theory (and numerous simulation studies) suggests that by improving the plausibility

of the MAR assumption, bias is likely to be reduced. Our proposed approach provides *one* potential solution to handling missing data, but is not a panacea. The sensitivity of analysis findings to the specific method used for missing data handling should be explored.

One variable for which this approach was not able to restore sample representativeness particularly well was the young person's sex. Although there was strong evidence of associations between this variable and most of the predictors of non-response at Wave 8, such associations with the Wave 1 auxiliary variables were not generally apparent. This is a clear difference for young person's sex relative to the associations seen for other variables, so forms a possible explanation for the observed sample representativeness results, indicating that more auxiliary variables associated with sex would be needed.

In our illustrative regression analysis, while the pattern of association was consistent across the CCA and MI analysis, the magnitude of association was generally somewhat greater in the latter. Although in this analysis there is no 'known truth' to use for comparison, the MAR assumption underlying the MI analysis would be considered far more plausible than the MCAR assumption underlying the CCA in this setting.

Existing literature

There has been recent interest in the representativeness of the respondents in Next Steps and how this may affect analyses of the data (Siddiqui et al, 2019). In this paper we have addressed the concerns raised by Siddiqui et al by proposing and demonstrating an approach whereby sample representativeness can be restored despite selective response. In particular, we have demonstrated that, using data from only Wave 8 respondents, through application of appropriate MI analyses we were largely able to restore the distributions of selected Wave 1 socio-demographic variables and Wave 1 and Wave 2 household salary, and could reliably estimate university attendance close to the population rate. Contrary to the assertions of Siddiqui et al, we therefore argue that Next Steps is a robust research resource for HE research, and indeed a range of other research uses.

Many of our identified predictors of non-response correspond to those previously identified in the literature on non-response in large British longitudinal surveys, including being male (Hawkes and Plewis, 2006; Atherton et al, 2008; Uhrig, 2008; Watson and Wooden, 2009; Lynn et al, 2012; Boyd et al, 2013; Mostafa and Wiggins, 2014; Fry et al, 2017; Lynn and Borkowska, 2018; Cornish et al, 2021; Mostafa et al, 2021), socio-economic disadvantage (Hawkes and Plewis, 2006; Plewis, 2007; Atherton et al, 2008; Plewis et al, 2008; Uhrig, 2008; Boyd et al, 2013; Lynn and Borkowska, 2018; Mostafa et al, 2021), changing address (Hawkes and Plewis, 2006; Plewis, 2007; Plewis et al, 2008; Uhrig, 2008; Mostafa, 2016), living in rented housing (Plewis, 2007; Atherton et al, 2008; Plewis et al, 2008; Uhrig, 2008; Lynn et al, 2012; Cornish et al, 2021), older age of main parent (usually mother) (Plewis, 2007; Mostafa and Wiggins, 2014; Cornish et al, 2021), childhood behavioural problems (Atherton et al, 2008; Mostafa et al, 2021), working or 'other' current main activity rather than full-time education (Uhrig, 2008), not consenting to data linkage with administrative records (Lagorio, 2016), and prior non-response (Watson and Wooden, 2014; Mostafa et al, 2021). Some additional predictors have only previously been identified in the broader non-response literature, such as being unable to access the

internet from home (Olson et al, 2012; Herzing and Blom, 2018) and smoking cigarettes (Kalsbeek et al, 2002; Cunradi et al, 2005; Young et al, 2006; McCoy et al, 2009).

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. Whether the young person is willing to answer questions on sexual experiences is essentially an issue of item non-response, which has been studied in relation to subsequent wave non-response (Loosveldt et al, 2002), though not considering questions of sexual experiences in particular. Some of these factors are quite specific, which may mean that they have not been considered as potential predictors of non-response in previous studies. While 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 these may differ across contexts and generations.

Strengths and limitations

There are many strengths to our study. We used 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. Similar data-driven approaches have been applied to non-response in different cohorts (Mostafa et al, 2021) and to administrative record linkage consent in the Next Steps cohort (Peycheva et al, 2021). 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 (Seaman et al, 2012) 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 could consider multilevel MI in this context (Quartagno et al, 2019).

As we used a multistage variable selection procedure, the final variance estimates (that is, 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 (Greenland, 2008), potentially leading to smaller p-values and hence false-positive inclusions within our ultimate set of predictors of non-response. However, since our $p < .001$ criterion is to some extent arbitrary, this is not a major limitation.

We were unable to exclude from the analysis cohort members who had died prior to Wave 8 or were no longer living in the UK and hence were no longer in the target population. However, we would expect the numbers of cohort members affected by this, particularly by mortality, to be low in this young cohort and therefore this is unlikely to make a meaningful difference to the findings of the study. Since we chose to combine the sequential events of contact and response within our single definition of non-response, we were not able to identify predictors of contact or response (given contact) individually.

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 HE institution 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 used since Wave 5 of Next Steps. Mode effects may plausibly have affected the values of the Wave 5–7 variables (de Leeuw, 2005; Goodman et al, 2022; Sakshaug et al, 2022), but given the strength of association required in the identification of predictors of 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.

MI is a powerful tool for addressing bias due to missing data, but care is required in its implementation. It is not the intention of this paper to provide a step-by-step guide to doing so – overview and guidance papers are available elsewhere (see, for example, Sterne et al, 2009; Azur et al, 2011; White et al, 2011) – but we emphasise here a few key aspects. It is vital that the imputation model is compatible with the analysis model: all variables in the analysis model must be present, including the outcome. If using chained equations, the individual regression models forming the imputation model must be correctly specified based on the nature and distributional features of each variable being imputed. The plausibility of the MAR assumption can be improved through inclusion of suitable auxiliary variables – those that are predictive of the underlying values of variables that are subject to missingness, and especially those that are also associated with the probability of data being missing, as have been identified here. Imputation models can be particularly prone to convergence issues, especially in situations with data sparsity. Convergence should be assessed by examining the individual regression models and through diagnostic approaches such as trace plots.

Future work

The present study focused on wave 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. 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 (Elliott and Shepherd, 2006; Sullivan et al, 2022) and the Millennium Cohort Study (Connelly and Platt, 2014; Joshi and Fitzsimons, 2016). Recent linkages of administrative data, including the

National Pupil Database (University College London et al, 2021) and Hospital Episode Statistics (University College London et al, 2020), into the Next Steps cohort data provide further information that may be of relevance to non-response. Subsequent work will therefore integrate such administrative data into the data-driven approach to the identification of predictors of non-response.

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, demonstrating that Next Steps is a robust data source for research. Our identification of these variables will allow users of the cohort to explore and attempt to reduce the 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.

Funding

This work was supported by the Economic and Social Research Council under Grant ES/M001660/1 and Grant ES/W013142/1.

Data availability statement

The authors take responsibility for the integrity of the data and the accuracy of the analysis. Next Steps data used in this paper are available from the UK Data Service (https://www.ukdataservice.ac.uk/).

Experimentation on humans and animals statement

Ethical approval for the Next Steps study was secured by the Centre for Longitudinal Studies from the NHS Research Ethics Committee (NRES) (REC Reference 14/LO/0096).The study complied with the principles of the Declaration of Helsinki. Study members participated on the basis of informed consent.

Conflict of interest

The authors declare that there is no conflict of interest.

References

Atherton, K., Fuller, E., Shepherd, P., Strachan, D.P. and Power, C. (2008) Loss and representativeness in a biomedical survey at age 45 years: 1958 British Birth Cohort, *Journal of Epidemiology and Community Health*, 62(3): 216–23. doi: 10.1136/jech.2006.058966

Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J. (2011) Multiple imputation by chained equations: what is it and how does it work?, *International Journal of Methods in Psychiatric Research*, 20(1): 40–9. doi: 10.1002/mpr.329

Boyd, A., Golding, J., Macleod, J., Lawlor, D.A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S. and Davey Smith, G. (2013) Cohort profile: the 'children of the 90s' –the index offspring of the avon longitudinal study of parents and children, *International Journal of Epidemiology*, 42(1): 111–27. doi: 10.1093/ije/dys064

Calderwood, L. (2018) *Next Steps: Sweep 8 – Age 25 Survey User Guide*, 2nd edn, London: UCL Centre for Longitudinal Studies.

Calderwood, L. and Sanchez, C. (2016) Next Steps (formerly known as the Longitudinal Study of Young People in England), *Open Health Data*, 4: art: e2. doi: 10.5334/ohd.16

Carpenter, J.R. and Kenward, M.G. (2013) *Multiple Imputation and Its Application*, Chichester: Wiley.

Connelly, R. and Platt, L. (2014) Cohort profile: UK Millennium Cohort Study (MCS), *International Journal of Epidemiology*, 43(6): 1719–25. doi: 10.1093/ije/dyu001

Cornish, R.P., Macleod, J., Boyd, A. and Tilling, K. (2021) Factors associated with participation over time in the Avon Longitudinal Study of Parents and Children: a study using linked education and primary care data, *International Journal of Epidemiology*, 50(1): 293–302. doi: 10.1093/ije/dyaa192

Cunradi, C., Moore, R., Killoran, M. and Ames, G. (2005) Survey nonresponse bias among young adults: the role of alcohol, tobacco, and drugs, *Substance Use & Misuse*, 40(2): 171–85. doi: 10.1081/ja-200048447

de Leeuw, E.D. (2005) To mix or not to mix data collection modes in surveys, *Journal of Official Statistics*, 21(2): 233–55.

DfE (Department for Education) (2011) Youth cohort study & longitudinal study of young people in england: the activities and experiences of 19 Year Olds – England 2010, Statistical Bulletin B01/2011, https://www.gov.uk/government/statistics/youth-cohort-study-and-longitudinal-study-of-young-people-in-england-the-activities-and-experiences-of-19-year-olds-2010.

DfE (Department for Education) (2018) Statistics: participation rates in higher education, https://web.archive.org/web/20190724032058/https://www.gov.uk/government/collections/statistics-on-higher-education-initial-participation-rates.

Elliott, J. and Shepherd, P. (2006) Cohort profile: 1970 British Birth Cohort (BCS70), *International Journal of Epidemiology*, 35(4): 836–43. doi: 10.1093/ije/dyl174

Enders, C.K. (2001) The performance of the full information maximum likelihood estimator in multiple regression models with missing data, *Educational and Psychological Measurement*, 61(5): 713–40. doi: 10.1177/00131640121971482

Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R. and Allen, N.E. (2017) Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population, *American Journal of Epidemiology*, 186(9): 1026–34. doi: 10.1093/aje/kwx246

Goodman, A., Brown, M., Silverwood, R.J., Sakshaug, J.W., Calderwood, L., Williams, J. and Ploubidis, G.B. (2022) The impact of using the web in a mixed mode follow-up of a longitudinal birth cohort study: evidence from the National Child Development Study, *Journal of the Royal Statistical Society: Series A – Statistics in Society*, 185(3): 822–50. doi: 10.1111/rssa.12786

Greenland, S. (2008) Invited commentary: variable selection versus shrinkage in the control of multiple confounders, *American Journal of Epidemiology*, 167(5): 523–9; discussion 530–1. doi: 10.1093/aje/kwm355

Harel, O., Mitchell, E.M., Perkins, N.J., Cole, S.R., Tchetgen Tchetgen, E.J., Sun, B.L. and Schisterman, E.F. (2018) Multiple imputation for incomplete data in epidemiologic studies, *American Journal of Epidemiology*, 187(3): 576–84. doi: 10.1093/aje/kwx349

Hawkes, D. and Plewis, I. (2006) Modelling non-response in the National Child Development Study, *Journal of the Royal Statistical Society: Series A – Statistics in Society*, 169(3): 479–91. doi: 10.1111/j.1467-985x.2006.00401.x

Herzing, J.M.E. and Blom, A.G. (2018) The influence of a person's digital affinity on unit nonresponse and attrition in an online panel, *Social Science Computer Review*, 37(3): 404–24. doi: 10.1177/0894439318774758

Joshi, H. and Fitzsimons, E. (2016) The Millennium Cohort Study: the making of a multi-purpose resource for social science and policy, *Longitudinal and Life Course Studies*, 7(4): 409–30. doi: 10.14301/llcs.v7i4.410

Kalsbeek, W.D., Yang, J. and Agans, R.P. (2002) *Predictors of Nonresponse in a Longitudinal Survey of Adolescents, ASA Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp 1740–5.

Lagorio, C. (2016) Call and Response: Modelling Longitudinal Contact and Cooperation Using Wave 1 Call Records Data, Understanding Society Working Paper No. 2016-01, Colchester, Institute for Social and Economic Research, University of Essex.

Little, R.J.A. and Rubin, D.B. (1989) The analysis of social-science data with missing values, *Sociological Methods & Research*, 18(2/3): 292–326. doi: 10.1177/0049124189018002004

Little, R.J.A. and Rubin, D.B. (2020) *Statistical Analysis with Missing Data*, 3rd edn, Hoboken, NJ: Wiley.

Loosveldt, G., Pickery, J. and Billiet, J. (2002) Item nonresponse as a predictor of unit nonresponse in a panel survey, *Journal of Official Statistics*, 18(4): 545–57.

Lynn, P. and Borkowska, M. (2018) Some Indicators of Sample Representativeness and Attrition Bias for BHPS and Understanding Society, Understanding Society Working Paper No. 2018-01, Colchester, Institute for Social and Economic Research, University of Essex.

Lynn, P., Burton, J., Kaminska, O., Knies, G. and Nandi, A. (2012) An Initial Look at Non-Response and Attrition in Understanding Society, Understanding Society Working Paper No. 2012-02, Colchester, Institute for Social and Economic Research, University of Essex.

McCoy, T.P., Ip, E.H., Blocker, J.N., Champion, H., Rhodes, S.D., Wagoner, K.G., Mitra, A. and Wolfson, M. (2009) Attrition bias in a U.S. internet survey of alcohol use among college freshmen, *Journal of Studies on Alcohol and Drugs*, 70(4): 606–14. doi: 10.15288/jsad.2009.70.606

Mostafa, T. (2016) Measuring the impact of residential mobility on response: evidence from the Millennium Cohort Study, *Longitudinal and Life Course Studies*, 7(3): 201–17. doi: 10.14301/llcs.v7i3.378

Mostafa, T. and Wiggins, R.D. (2014) Handling Attrition and Non-Response in the 1970 British Cohort Study, CLS Working Paper 2014/2, London, UCL Centre for Longitudinal Studies.

Mostafa, T., Narayanan, M., Pongiglione, B., Dodgeon, B., Goodman, A., Silverwood, R.J. and Ploubidis, G.B. (2021) Missing at random assumption made more plausible: evidence from the 1958 British Birth Cohort, *Journal of Clinical Epidemiology*, 136: 44–54. doi: 10.1016/j.jclinepi.2021.02.019

Olson, K., Smyth, J.D. and Wood, H.M. (2012) Does giving people their preferred survey mode actually increase survey participation rates? An experimental examination, *Public Opinion Quarterly*, 76(4): 611–35. doi: 10.1093/poq/nfs024

Peycheva, D., Ploubidis, G. and Calderwood, L. (2021) Determinants of consent to administrative records linkage in longitudinal surveys: evidence from next steps, in P. Lynn (ed) *Advances in Longitudinal Survey Methodology*, Chichester: Wiley, pp 151–80.

Plewis, I. (2007) Non-response in a birth cohort study: the case of the Millennium Cohort Study, *International Journal of Social Research Methodology*, 10(5): 325–34. doi: 10.1080/13645570701676955

Plewis, I., Ketende, S.C., Joshi, H. and Hughes, G. (2008) 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*, 24(3): 365–85.

Quartagno, M., Carpenter, J.R. and Goldstein, H. (2019) Multiple imputation with survey weights: a multilevel approach, *Journal of Survey Statistics and Methodology*, 8(5): 965–89. doi: 10.1093/jssam/smz036

Rubin, D.B. (1976) Inference and missing data, *Biometrika*, 63(3): 581–92. doi: 10.1093/biomet/63.3.581

Rubin, D.B. (2004) *Multiple Imputation for Nonresponse in Surveys*, Hoboken, NJ: Wiley.

Sakshaug, J.W., Cernat, A., Silverwood, R.J., Calderwood, L., Ploubidis, G.B. and Goodman, A. (2022) Measurement equivalence in sequential mixed-mode surveys: evidence from the next steps cohort study, *Survey Research Methods*, 16(1): 29–43. doi: 10.18148/srm/2022.v16i1.7811

Schafer, J.L. and Olsen, M.K. (1998) Multiple imputation for multivariate missing-data problems: a data analyst's perspective, *Multivariate Behavioral Research*, 33(4): 545–71. doi: 10.1207/s15327906mbr3304_5

Seaman, S.R. and White, I.R. (2013) Review of inverse probability weighting for dealing with missing data, *Statistical Methods in Medical Research*, 22(3): 278–95. doi: 10.1177/0962280210395740

Seaman, S.R., White, I.R., Copas, A.J. and Li, L. (2012) Combining multiple imputation and inverse-probability weighting, *Biometrics*, 68(1): 129–37. doi: 10.1111/j.1541-0420.2011.01666.x

Siddiqui, N., Boliver, V. and Gorard, S. (2019) Reliability of longitudinal social surveys of access to higher education: the case of Next Steps in England, *Social Inclusion*, 7(1): 80–9. doi: 10.17645/si.v7i1.1631

Spratt, M., Carpenter, J., Sterne, J.A.C., Carlin, J.B., Heron, J., Henderson, J. and Tilling, K. (2010) Strategies for multiple imputation in longitudinal studies, *American Journal of Epidemiology*, 172(4): 478–87. doi: 10.1093/aje/kwq137

StataCorp (2017) Stata Statistical Software: Release 17, College Station, TX: StataCorp LLC.

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M. and Carpenter, J.R. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, *BMJ*, 338: art b2393. doi: 10.1136/bmj.b2393

Sullivan, A., Brown, M., Hamer, M. and Ploubidis, G.B. (2022) Cohort profile update: the 1970 British Cohort Study (BCS70), *International Journal of Epidemiology*, 52(3): e179–86. doi: 10.1093/ije/dyac148

Uhrig, S.C.N. (2008) The Nature and Causes of Attrition in the British Household Panel Study, ISER Working Paper No. 2008-05, Colchester, Institute for Social and Economic Research, University of Essex.

University College London, UCL Institute of Education and Centre for Longitudinal Studies (2021) *Next Steps: Linked Education Administrative Datasets (National Pupil Database), England, 2005–2009: Secure Access SN: 7104 [data collection]*, 6th edn, Colchester: UK Data Service, https://beta.ukdataservice.ac.uk/datacatalogue/doi/?id=7104#!#4.

University College London, UCL Institute of Education, Centre for Longitudinal Studies and NHS Digital (2020) *Next Steps: Linked Health Administrative Datasets (Hospital Episode Statistics), England, 1997–2017: Secure Access SN: 8681 [data collection]*, Colchester: UK Data Service, https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8681.

University of London, UCL Institute of Education and Centre For Longitudinal Studies (2018) *Next Steps: Sweeps 1–8, 2004–2016 SN: 5545 [data collection]*, 14th edn, Colchester: UK Data Service, https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=5545&type=Data%20catalogue&lt.

Watson, N. and Wooden, M. (2009) Identifying factors affecting longitudinal survey response, in P. Lynn (ed) *Methodology of Longitudinal Surveys*, Chichester: Wiley, pp 157–82.

Watson, N. and Wooden, M. (2014) Re-engaging with survey non-respondents: evidence from three household panels, *Journal of the Royal Statistical Society: Series A – Statistics in Society*, 177(2): 499–522. doi: 10.1111/rssa.12024

White, I.R. and Carlin, J.B. (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values, *Statistics in Medicine*, 29(28): 2920–31. doi: 10.1002/sim.3944

White, I.R., Royston, P. and Wood, A.M. (2011) Multiple imputation using chained equations: issues and guidance for practice, *Statistics in Medicine*, 30(4): 377–99. doi: 10.1002/sim.4067

Wooldridge, J.M. (2007) Inverse probability weighted estimation for general missing data problems, *Journal of Econometrics*, 141(2): 1281–301. doi: 10.1016/j.jeconom.2007.02.002

Young, A.F., Powers, J.R. and Bell, S.L. (2006) Attrition in longitudinal studies: who do you lose?, *Australian and New Zealand Journal of Public Health*, 30(4): 353–61. doi: 10.1111/j.1467-842x.2006.tb00849.x

Zou, G. (2004) A modified Poisson regression approach to prospective studies with binary data, *American Journal of Epidemiology*, 159(7): 702–6. doi: 10.1093/aje/kwh090