The impact of attrition and non-response in birth cohort studies: a need to incorporate missingness strategies

Dick Wiggins , Quantitative Social Science and

Tarek Mostafa, Centre for Longitudinal Studies,

Department of Social Science

Challenges to Longitudinal Surveys

- Statistical analyses face a number of challenges:
- Unit non-response.
- Item non-response.
- Attrition over time in longitudinal surveys.

- *If analysts ignore these challenges our analysis will be biased and under-powered.*

Challenges to Longitudinal Surveys

Dual Aim of the paper:

- To raise user's appreciation of the need to incorporate strategies to handle missingness in any longitudinal analysis of the 1970 British Cohort Study (BCS70) and, indeed birth cohort studies more generally.

- To illustrate and (evaluate) the use of non-response weights and multiple imputation to deal with attrition and item missingness

British Cohort Study 1970

BCS70 follows the lives of 17,000 people born in a single week in April 1970. It is an extensive multipurpose longitudinal survey which collects individual data on health, physical, educational and social development, and economic circumstances among other areas.

Sweep	Age (years)	Year
1	Birth	1970
2	5	1975
3	10	1980
4	16	1986
5	26	1996
6	30	2000
7	34	2004
8	38	2008
9	42	2012

Defining Attrition

Attrition is the discontinued participation of some individuals in a longitudinal survey for reasons that are unknown and/or beyond the control of the researcher

Patterns of response

Pattern	Type of non-response
111111111	Participated in all sweeps
111111110	Monotone response
111100000	Monotone response
111101111	Non-monotone response
100000011	Non-monotone response

Pattern	Frequency	Percentage
Monotone	4,716	27
Non monotone	9,153	53
Participated in all sweeps	3,423	20
Total	17,284	100

A little theoretical background

- Historically, we encounter two broad approaches which have been adopted under MAR to handle the problem of missing data (see Kalton (1986) and Lepowski (1989))

- The application of Inverse Probability Weights (IPWs) and/or

- Multiple Imputation (MI)

- A combination of the use of both IPWs and MI

A little theoretical background continued

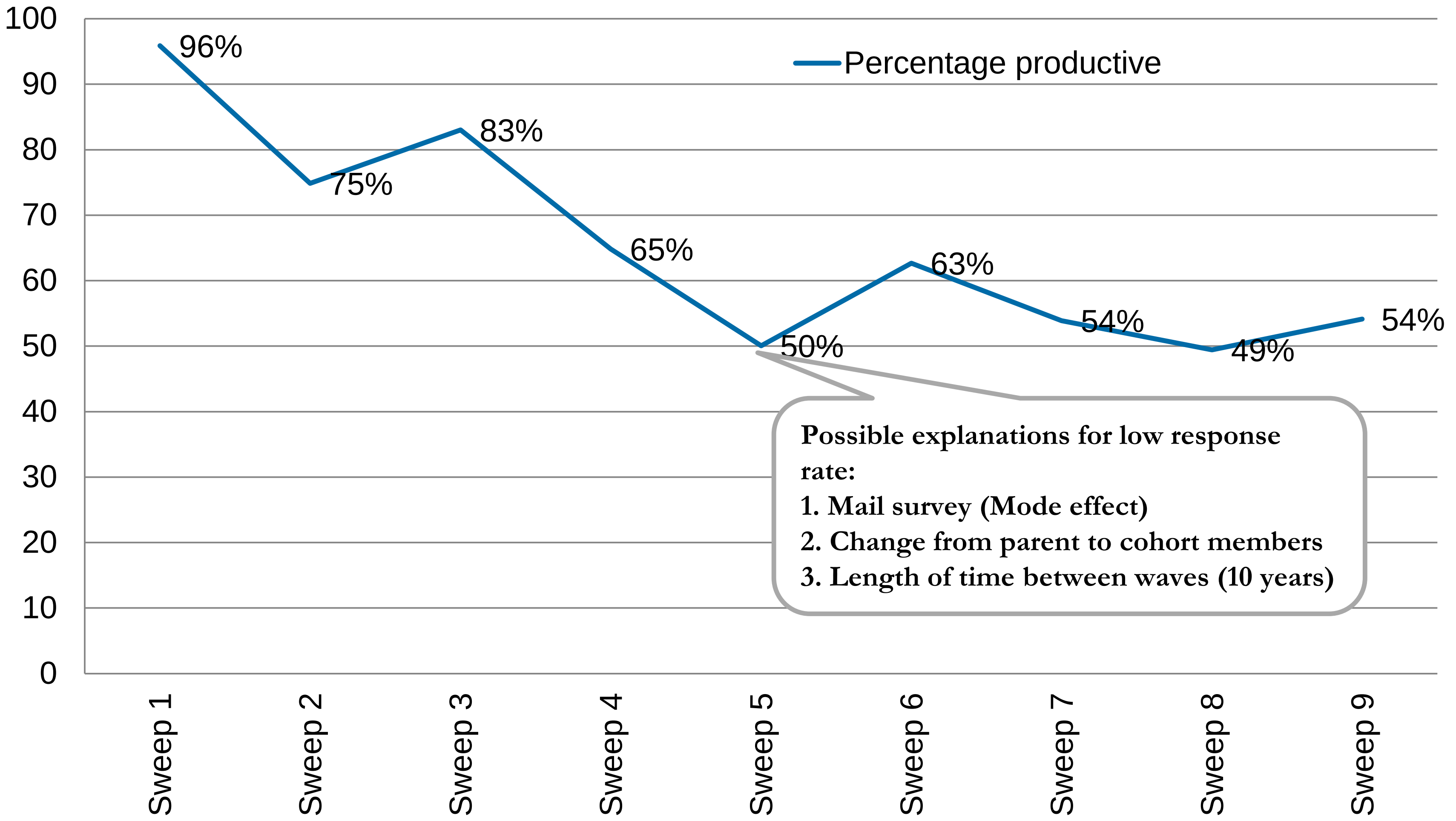
- Tackling the problem of 'information loss' requires the user to make assumptions about the 'missingness mechanism' Following (Rubin (1976), Little & Rubin (2002), Carpenter & Plewis (2011) and others these mechanisms are described as

- Missing Completely at Random (MCAR)

- Missing At Random (MAR)

- Missing Not at Random (MNAR)

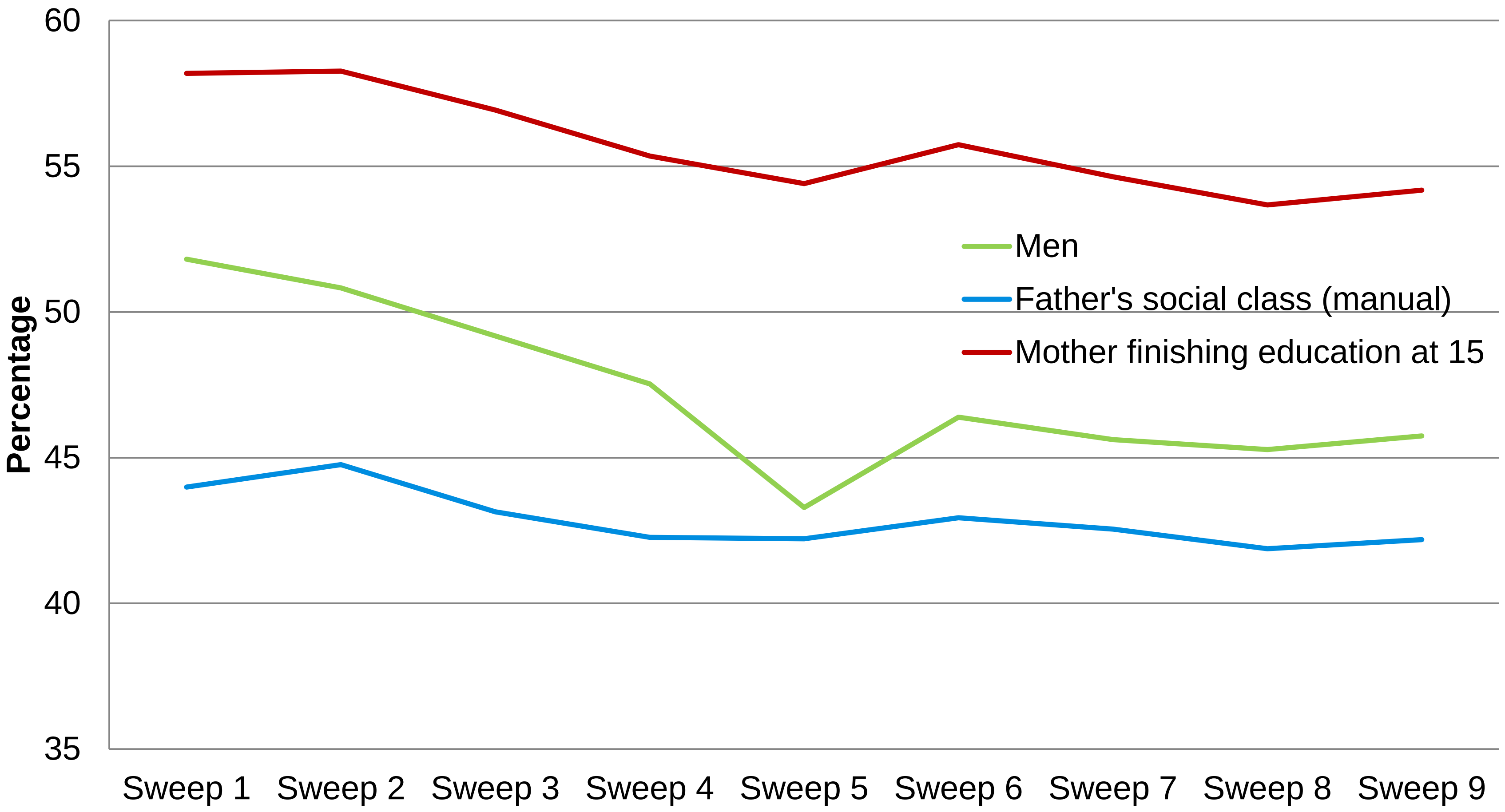
Overall response rates for BCS70



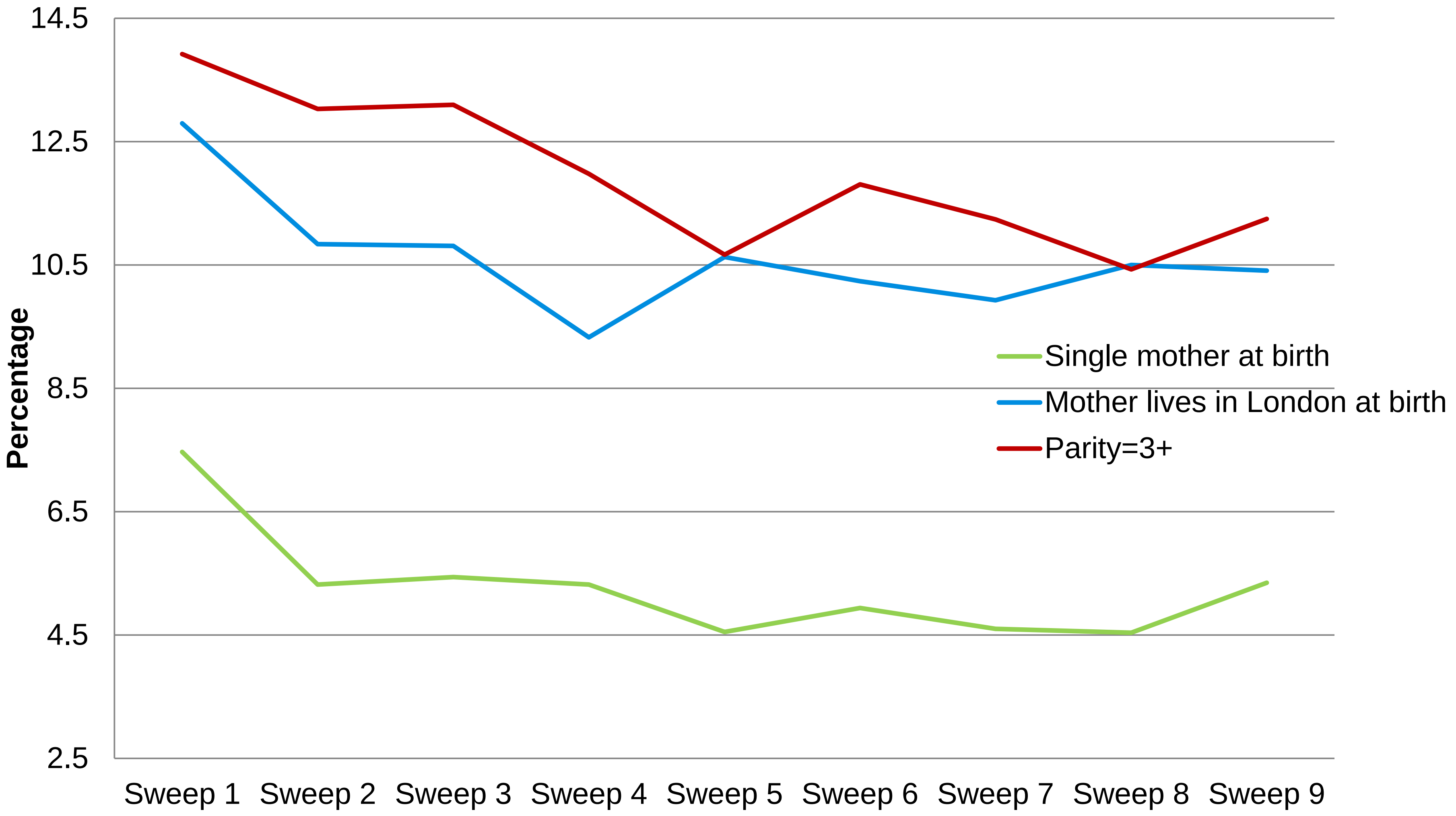
Changing sample composition over time according to a cohort member's parental characteristics at birth

- **Gender**
- **Dad's social class (% manual)**
- **Mum' finishing full-time education at 15 years**
- **Mum single at birth of CM**
- **Mum lived in London at birth of CM**
- **Mum's parity > 3**

Sample Composition Over Time



Sample Composition Over Time



Modelling response in terms of a CM’s birth characteristics

- Applied a series of binary logistic regression models for each sweep of data collection
- Caution: resulting ‘goodness of fit’ is poor (pseudo-R^2 varies between 2.5 and 3.6 per cent)
- In sum: women more likely to respond than men, CMs born to a single Mum less likely to respond as do CMs whose mother’s didn’t attempt to breast feed or had older siblings. Response probabilities also increase with age of Mum at birth and for Mum’s with a longer formal education.
- Father’s social class and age at completing full-time education were also important.

Construction IPWs on the basis of modelling response in terms of a CM's birth characteristics

- 'Weights' recover selection bias, for example consider the percentage estimate of Mum's who live in London at time of CM's birth by sweep 4:

Sweep	Percentage estimate
1 (at birth)	12.3
4 without reweighting	9.5
4 with reweighting	12.3

Multiple Imputation (MI)

- Under MI (Little & Rubin, 2002), Schafer & Olsen (1998) and Rubin (1987, 2004) missing values under MAR assumptions are replaced several times to create filled-in replicates of our data. These replicates are analysed separately and ultimately combined under Rubin’s Rules (Rubin, 1987).

- Approaches to MI vary which vary according to the type of data to be imputed. For instance the level of measurement and/or the data structure.

- Recommend Carpenter & Kenward (2013) for a valuable overview.

How effective are weights and imputations?

To recap 'two heads are better then one' !!

Using a combination of :

- Attrition (inverse probability) weights alone have: weak predictive power, no solution to item missingness, constructed using restrictive models, reduction in sample size especially when using data from different waves.
- Multiple Imputation: can be used to treat both unit and item non-response, can be tailored according to the need of the researcher.
- Both techniques require knowledge of the process behind missingness.

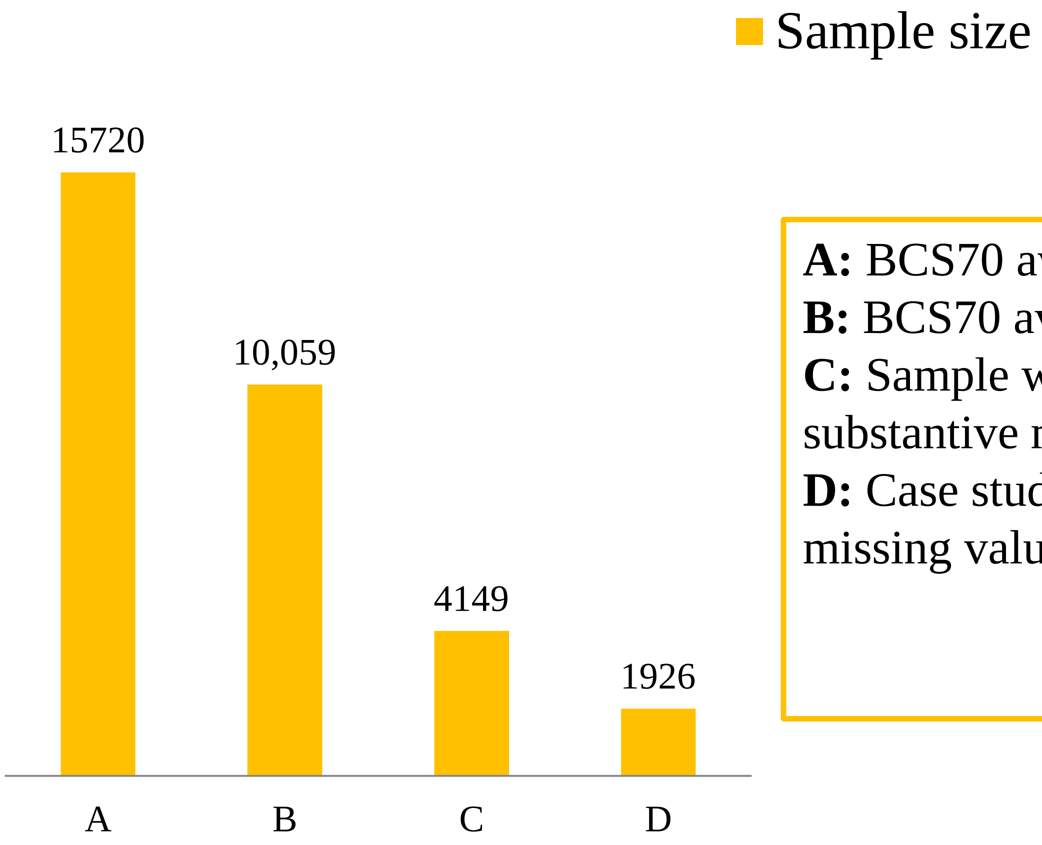
Single Case Study illustration

- Use a hypothetical substantive model with
- **Dependent variable**: vocabulary scores at age 16 (Parsons, 2014)
- **Independent variables**: gender, age 10 gross family income per week (sweep 3) and highest parental qualification (sweep 4).
- **Case study creation**:
- 1- Construct inverse probability weights for sweep 4. These weights will adjust for attrition in wave 4.
- 2- On literacy scores, introduce 10% missing values completely at random.
- 3- We recode the father’s social class into a binary variable with two categories manual and non-manual.
- 4- On income and highest qualification, we introduce 40% missing values if the father’s social class is manual and 10% if it is non-manual.
- 5- We don’t introduce any missing values on gender.

Case Study illustration continued – the sample stages

Sample label	Sample Size	Comment
A	15,720	Sweep 1 potential
B	10,059	Sweep 4
C	4,149	Complete cases given our substantive model
C_{miss}	4,149	C with missing data imposed
D	1,926	Complete cases based on listwise deletion for C_{miss}

Case study stages



A: BCS70 available sample size at sweep 1.
B: BCS70 available sample size at sweep 4.
C: Sample with complete cases given the substantive model.
D: Case study sample after the introduction of missing values.

Modelling strategy

Models only vary according to the use of Inverse Probability Weights (IPWs) and/or

Multiple Imputations (MI) to fill-in

Model	IPW adjustment	MI to fill-in (20 replicates)
1. “Benchmarker C”	Yes	No since complete
2. As for model 1	No	No
3. “Days gone by D”	No	No since reduced C_{miss}
4. As for model 3 but with	Yes	ditto
5. 1990’s model	No	Yes MI on C_{miss}
6. “preferred model”	Yes	ditto

MI procedure in STATA

- For models 5 & 6 use a Markov-Chain-Monte-Carlo procedure (Gilks, Richardson & Spiegelhalter (1996) and Chained equations (Royston, 2009 also Royston & White, 2011).
- In STATA use MI command with a linear procedure to impute vocabulary scores (continuous) and ordinal-logit for income and highest educational qualification.
- Auxiliary variables used in the imputation model were birth characteristics (parental marital status, parity, whether or not Mum attempted to breast feed, mother's age at delivery, mother and father's age at completing full-time education). Dad's social class not included as this variable was used to introduce item missingness.

What do the models convey?

- Models 3 and 4 are the least reliable in terms of parameter estimates and size of standard errors

- Models 5 and 6 are closest to model 1 apart from the estimates effects of a parent's highest educational qualification

- We return a 'mixed picture'

Modelling results

Model	IPW adjustment	MI to fill-in (20 replicates)	In relation to benchmarker model
1. “Benchmarker C”	Yes	No since complete	NA
2. As for model 1 but no IPW adjustment	No	No	Close apart from income effect
3. “Days gone by D”	No	No since reduced C_{miss}	Unreliable
4. As for model 3 but with	Yes	ditto	Similar to model 3 - unreliable
5. 1990’s model	No	Yes MI on C_{miss}	Close to benchmark apart from highest qualification effect
6. “preferred model”	Yes	ditto	As above

Conclusions

- The analysis of response rates reveal some interesting patterns of non-response based on a CM's birth characteristics
- But the predictive power of these models is weak.
- Non-response weights don't improve the estimates or their standard errors by much when data loss is due to a selected pattern of item missingness.
- Multiple imputations are efficient in reducing the bias resulting from item missingness both in terms of parameter estimates and standard errors but with some worrying exceptions in our illustration.
- The efficacy of weights and imputations in dealing with bias resulting from unit non-response and item missingness depends on the extent of bias and whether variables correlated with the probability of unit and item non-response can be found.

Next steps

- Full simulation model with varying degrees of item missingness

- Sensitivity analysis: ‘joint modelling’ aka ‘Heckman modelling’ (1979) where the substantive model of interest is modelled jointly with a model for missingness. In this way it is proposed that the unobserved variables that simultaneously influence both the outcome and the missingness are captured in the residuals of the two models

- Technically, the challenge is to identify variables (or instruments) for the missingness model which predict the probability of missingness but do not correlate with the substantive outcome. See Carpenter & Plewis (2011) for an illustration using NCDS data.

Talk based upon

Mostafa, T. and Wiggins, R.D. (2015).The impact of attrition and non-response in birth cohort studies: a need to incorporate missingness strategies. Longitudinal and Life Course Studies Vol. 6, Issue No.2 pp 131-146.

Thank you for your attention.

r.wiggins@ioe.ac.uk

t.mostafa@ioe.ac.uk

Complimentary slides

- **Response categories over time**
- **Modelling non-response for successive sweeps in terms of birth characteristics**
- **Applying (hypothetical) regression models for various strategies to handle attrition and/or item missingness**

Response Categories

Response categories	Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8	Sweep 9
Participated	16,569	12,939	14,349	11,206	8,654	10,833	9,316	8,545	9,354
Dead	0	565	585	597	697	748	795	824	853
Unproductive	715	3,780	2,350	5,481	7,933	5,703	7,173	7,915	7,077
Total	17,284	17,284	17,284	17,284	17,284	17,284	17,284	17,284	17,284

Modelling Non-response in BCS70

	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8	Sweep 9
Gender (reference: Men)								
Women	1.00	1.08	1.26***	1.80***	1.49***	1.48***	1.48***	1.44***
	(0.040)	(0.049)	(0.044)	(0.060)	(0.052)	(0.049)	(0.049)	(0.048)
Marital status (reference: Single)								
Married	1.47***	2.18***	1.67***	1.85***	1.89***	1.89***	1.79***	1.42***
	(0.140)	(0.218)	(0.151)	(0.174)	(0.171)	(0.174)	(0.169)	(0.128)
Mother lives in London in 1970 (reference: not in London)								
In London	0.57***	0.55***	0.47***	0.71***	0.61***	0.62***	0.70***	0.67***
	(0.032)	(0.034)	(0.024)	(0.037)	(0.031)	(0.032)	(0.036)	(0.034)
Parity (reference: 0)								
1	0.97	1.02	0.87**	0.92	0.94	0.89**	0.93	0.92*
	(0.050)	(0.059)	(0.039)	(0.039)	(0.042)	(0.038)	(0.039)	(0.039)
2	0.82**	0.89	0.81***	0.79***	0.84**	0.74***	0.75***	0.81***
	(0.053)	(0.065)	(0.046)	(0.042)	(0.047)	(0.040)	(0.040)	(0.044)
3+	0.72***	0.90	0.70***	0.58***	0.65***	0.58***	0.54***	0.61***
	(0.053)	(0.076)	(0.045)	(0.036)	(0.041)	(0.036)	(0.033)	(0.038)

Modelling Non-response in BCS70

	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8	Sweep 9
Breastfeeding (reference: attempted)								
Not attempted	0.82***	0.84***	0.85***	0.85***	0.92*	0.87***	0.87***	0.80***
	(0.036)	(0.041)	(0.032)	(0.031)	(0.034)	(0.031)	(0.031)	(0.029)
Mother's age at Delivery (reference: less than 20)								
[20-24]	1.42***	1.17	1.20**	1.31***	1.23**	1.33***	1.28***	1.26***
	(0.105)	(0.098)	(0.080)	(0.085)	(0.080)	(0.085)	(0.083)	(0.081)
[25-29]	1.51***	1.27**	1.28***	1.46***	1.35***	1.50***	1.45***	1.35***
	(0.121)	(0.115)	(0.092)	(0.102)	(0.096)	(0.103)	(0.101)	(0.093)
[30-34]	1.63***	1.36**	1.30**	1.62***	1.44***	1.66***	1.59***	1.39***
	(0.151)	(0.143)	(0.106)	(0.129)	(0.117)	(0.131)	(0.125)	(0.109)
35 or more	1.81***	1.56***	1.40***	1.69***	1.51***	1.81***	1.73***	1.45***
	(0.204)	(0.198)	(0.140)	(0.164)	(0.149)	(0.175)	(0.167)	(0.139)
Mother's age at completion of education (reference: 14 or less)								
15	1.56***	1.81***	1.29**	1.38***	1.20*	1.32***	1.15	1.04
	(0.141)	(0.179)	(0.106)	(0.114)	(0.098)	(0.107)	(0.094)	(0.084)
16	1.63***	1.73***	1.50***	1.50***	1.37***	1.51***	1.34***	1.21*
	(0.164)	(0.190)	(0.137)	(0.135)	(0.124)	(0.134)	(0.119)	(0.107)
17	1.47***	1.42**	1.32**	1.56***	1.26*	1.45***	1.32**	1.18
	(0.172)	(0.180)	(0.138)	(0.160)	(0.131)	(0.148)	(0.134)	(0.120)
18 or more	1.31*	1.34*	1.30*	1.48***	1.14	1.33**	1.24*	1.05
	(0.147)	(0.164)	(0.133)	(0.149)	(0.116)	(0.134)	(0.124)	(0.105)

Modelling Non-response in BCS70

	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8	Sweep 9
Father's social class (reference: SC 1)								
Professional	0.94	0.98	0.85	0.94	0.93	0.99	0.95	0.97
	(0.102)	(0.116)	(0.084)	(0.087)	(0.090)	(0.092)	(0.088)	(0.090)
Clerical, non-manual	1.06	1.20	1.04	1.07	1.10	1.13	0.99	1.00
	(0.122)	(0.151)	(0.107)	(0.102)	(0.111)	(0.108)	(0.094)	(0.095)
Skilled manual	0.90	0.94	0.79*	0.79*	0.83*	0.84	0.74***	0.77**
	(0.097)	(0.109)	(0.076)	(0.071)	(0.078)	(0.076)	(0.067)	(0.070)
Unskilled manual	0.87	0.85	0.75**	0.70***	0.76**	0.75**	0.68***	0.69***
	(0.101)	(0.108)	(0.079)	(0.068)	(0.077)	(0.073)	(0.066)	(0.068)
Lowest grade workers	0.70**	0.77	0.69**	0.56***	0.64***	0.65***	0.56***	0.59***
	(0.091)	(0.111)	(0.081)	(0.063)	(0.074)	(0.072)	(0.063)	(0.065)
Other	0.34***	0.60***	0.70**	0.70**	0.69**	0.76*	0.65***	0.70**
	(0.044)	(0.085)	(0.086)	(0.082)	(0.083)	(0.088)	(0.075)	(0.081)

Modelling Non-response in BCS70

	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8	Sweep 9
Father's age at completion of education (reference: 14 or less)								
15	1.20*	1.24*	1.11	1.02	1.19*	1.03	1.11	1.03
	(0.102)	(0.119)	(0.083)	(0.076)	(0.089)	(0.076)	(0.082)	(0.076)
16	1.09	1.00	1.14	1.07	1.13	1.00	1.10	0.99
	(0.107)	(0.108)	(0.098)	(0.090)	(0.096)	(0.084)	(0.092)	(0.082)
17	0.92	1.04	1.25*	1.21	1.27*	1.10	1.29*	1.08
	(0.107)	(0.136)	(0.131)	(0.122)	(0.132)	(0.111)	(0.130)	(0.108)
18 or more	0.79*	0.82	0.98	0.96	1.05	1.00	1.06	0.92
	(0.083)	(0.094)	(0.092)	(0.088)	(0.097)	(0.091)	(0.097)	(0.083)
N	15270	15270	15270	15270	15270	15270	15270	15270
pseudo R^2	0.036	0.034	0.026	0.040	0.028	0.031	0.033	0.025

Results

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Gender						
Women	8.99**	8.41**	5.50	5.93	9.91**	10.4***
	(2.921)	(2.921)	(4.256)	(4.248)	(3.098)	(3.099)
Age 10 gross family income per week (reference: under £50)						
£50 - £99	3.61	2.22	0.90	2.93	1.97	2.78
	(7.464)	(7.538)	(12.054)	(11.919)	(9.508)	(9.532)
£100 - £149	9.40	7.64	2.24	4.45	8.19	8.92
	(7.443)	(7.507)	(11.994)	(11.865)	(8.989)	(9.132)
£150 - £199	14.4	12.9	5.54	7.54	10.0	10.6
	(7.941)	(7.989)	(12.641)	(12.536)	(9.913)	(10.073)
£200 - £249	13.7	11.9	10.5	12.8	17.7	18.7
	(9.184)	(9.195)	(13.898)	(13.820)	(11.240)	(11.468)
£250 or more	28.2**	27.2**	23.4	26.3	27.0*	28.0*
	(9.523)	(9.502)	(14.131)	(14.078)	(11.140)	(11.478)

Results

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Parental highest qualification (reference: no qualification)						
Other	24.3*	26.4*	30.6	29.9	20.7	19.8
	(11.816)	(11.897)	(16.687)	(16.410)	(13.989)	(13.807)
Vocational	16.9***	17.2***	23.4**	23.6***	13.8**	14.2**
	(4.502)	(4.545)	(7.256)	(7.163)	(5.262)	(5.265)
O level	32.8***	34.1***	40.1***	38.9***	28.4***	27.9***
	(4.250)	(4.246)	(6.485)	(6.465)	(5.286)	(5.235)
A level	51.3***	51.5***	56.9***	57.3***	44.1***	44.5***
	(5.533)	(5.477)	(8.135)	(8.153)	(7.180)	(7.320)
Nurse	54.5***	56.5***	46.3**	43.1**	47.6***	46.3***
	(9.536)	(9.381)	(14.076)	(14.272)	(10.854)	(11.129)
Teacher	70.3***	69.1***	74.6***	75.4***	59.1***	60.1***
	(9.115)	(8.994)	(12.006)	(12.071)	(9.470)	(9.563)
Higher degree	85.6***	84.5***	90.7***	91.1***	73.1***	74.0***
	(5.288)	(5.250)	(7.434)	(7.444)	(6.000)	(6.030)
Constant	-55.3***	-52.1***	-44.4***	-48.0***	-49.6***	-52.3***
	(8.428)	(8.517)	(13.288)	(13.142)	(9.940)	(9.885)
N	4149	4149	1926	1926	4149	4149

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$