SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

RS7: Data Integration

Prof Alexandru Cernat (RS lead, Uni. Manchester)
Prof Natalie Shlomo (Uni. Manchester)
Prof Nikos Tzavidis (Uni. Southampton)
Prof Joseph Sakshaug (IAB)

Dr Thomas O'Toole (Uni. Manchester)
thomas.otoole@manchester.ac.uk

University of Essex
University of Southampton
UKRI Economic and Social Research Council

NCRM National Centre for Research Methods
Office for National Statistics
UCL
Warwick The University of Warwick
Manchester 1824 The University of Manchester
City University of London Est 1894
National Centre for Social Research
LSE The London School of Economics and Political Science
Ulster University
UNIL | Université de Lausanne
Ipsos
Kantar Public

What is Data Integration?

- Data integration refers to the process of **bringing together information from multiple data sources in a coherent and consistent manner.**
  - Data integration makes it possible to examine relationships between factors which might not be visible from any one data source alone.

- ***Research strand 7 of Survey Futures is concerned with why and how non-survey data can be used to enhance survey data.***

What are our Research Themes?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Practice Guide 1	Practice Guide 2	Practice Guide 3
Options for integrating non-survey and population survey data.	**Using integrated non-survey data to evaluate and compensate for non-response bias in surveys.**	**Using integrated non-survey data for monitoring and intervening in survey data collection.**

What are our Research Themes?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Practice Guide 1 Recap

Options for integrating non-survey and population survey data.

https://surveyfutures.net/practice-guides/

University of Essex
University of Southampton
UKRI Economic and Social Research Council

Integrated non-survey data

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Administrative data

Administrative data is primarily collected for **routine, operational purposes**, and is recorded when an individual interacts with a service (Harron et al., 2017).

For example:
- **Health data**
- **Education data**
- **Employment and income data**

Administrative data is often linked to survey data at the **individual level.**

Geospatial data

Geospatial data is collected via satellite imagery or sensors.

For example:
- ***Government region***
- ***Middle/Lower Super Output Area (M/LSOA)***
- ***Postcode***
- ***km x km grid***
- ***Respondent unit***

These variables can be linked at the **selected spatial scale** to add contextual geospatial variables for each respondent.

Digital trace data

Digital trace data is derived from interactions with digital platforms, capturing real-time behaviours and trends (Boeschoten et al., 2022).

For example:
- **Web scraping**
- **Smart apps**
- **Document scanning**
- **Data donation**

Digital trace data is often linked at the individual-level and **collected/donated by survey respondents.**

How are Data Sources Integrated?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Deterministic Matching

- Records can be matched using an exact matching procedure **(i.e. National Insurance Number; NINO).**
- Or on a series of non-unique identifiers and multiple respondent characteristics, such as **sex and date of birth.**
- ‘**Fuzzy**’ matching allows for some errors in the identifiers.
- May lead to a higher rate of **false negatives** (or missed matches; **Harron et al., 2017).**

=

Probabilistic Matching

- Uses statistical modelling to obtain the **probability of a correct match (Fellegi and Sunter, 1969).**
- Probabilistic matching can be used when the criteria for **deterministic matching cannot be met** exactly.
- This method of data linkage may lead to a higher rate of **false positives** (or identified non-matches; **Harron et al., 2017).**

~

Accessing Integrated Data

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- In the United Kingdom, survey-to-non-survey integrated data is often available via the data holder's secure access service.
- Compulsory accredited researcher status under the Digital Economy Act (2017).
- Outputs must adhere to ethical and statistical disclosure requirements.

The most prominent secure services (or safe havens) include:

- ***The UK Data Service (UKDS)***
- ***The Office for National Statistics (ONS)***
- ***The UK Longitudinal Linkage Collaboration (UKLLC)***
- ***The Secure Anonymised Information Linkage (SAIL) Databank***
- ***Research Data Scotland's (RDS) Research Access Service***

What are our Research Themes?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Practice Guide 2

Using integrated non-survey data to evaluate and compensate for non-response bias in surveys.

1. What is non-response bias?
2. Methods for evaluating and compensating for non-response bias
3. Additional data sources for handling non-response bias
4. Recommendations and summary

What is survey non-response?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Over the past decade, response rates in probability-based surveys have seen a steady decline.
  - *The Labour Force Survey recorded a decline from 45 per cent in 2015 to just 17 per cent in 2024* ***(Office for National Statistics, 2015; 2024).***
- The Covid-19 pandemic amplified existing issues in survey non-response and created several new problems.
- However, a low response rate does not necessarily indicate poor survey quality **(Groves, 2006).**
- Survey non-response can be a useful tool, when used in addition to indicators of sample representativeness and composition **(Maslovskya et al., 2025).**

How does non-response occur?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- In **cross-sectional** surveys, unit (individual) non-response comes from a failure to successfully recruit a unit of interest.

- In **longitudinal** and **panel** surveys, non-response can also occur when units from the previous sweep of data collection are unable to be observed in the next sweep.

- This typically stems from **cumulative** survey attrition, as initially willing participants drop out of the study at later waves (Sakshaug, 2022).

How does non-response occur?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- A low response rate means that only a subset of the selected probability sample is ever measured, who *may not represent* who the target sample (Groves & Peytcheva, 2008).
- **Non-response is the property of a survey, non-response bias is the property of a statistic** (Wagner, 2012).

- The **probability** of non-response is **conditional** on **separate, common and survey** factors (Groves & Peytcheva, 2008).
- **Non-response bias** occurs when these causes result in an achieved sample which is **systematically different** from those who are missing.
  - *If there is a **relationship** between the variable response propensity and the survey variable.*

What is missing data?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Missing data in general can be classified according to Rubin (1976) in three mechanisms:

Missing Completely at Random (MCAR)	Missing at Random (MAR)	Missing Not at Random (MNAR)
*The probability of missing data is **independent of any observed or unobserved data.***	*The probability of missing data is **related to the observed data.***	*The probability of missing data is **related to the unobserved data.***

Through data integration, researchers can shift the mechanism of missingness from MNAR to MAR by integrating non-survey data to explain causes of non-response and non-linkage.

Evaluating non-response bias

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Population total comparisons

Compare survey estimates of demographic characteristics to known population totals from external sources, such as censuses or administrative records (Skalland, 2011).

Discrepancies between survey estimates and population totals can indicate coverage errors or non-response bias, prompting adjustments through poststratification.

Sub-group comparisons

Calculate response rates within specific demographic or other subgroups to identify patterns of non-response, typically by comparison to survey or population totals (Peycheva & Groves, 2008).

By examining these rates, researchers can detect whether certain subgroups are underrepresented, which may indicate potential non-response bias.

Evaluating non-response bias

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

R-indicators

Measure the **representativeness** of the survey response by quantifying the variation in response propensities across the sample (Bethlehem, Cobben &, Schouten, 2008; Plewis & Shlomo, 2017).

$$R = 1 - 2sd(p)$$

Higher values indicate low variability.

Can be extended with **distance measures** to quantify the **difference** between the sample and benchmark (ONS, 2023).

Coefficients of variation

Measure of the **relative variability** of the response propensities in the responding sample (Moore, Durrant & Smith, 2018; Schouten & Shlomo, 2017).

$$CV = \frac{sd(p)}{p}$$

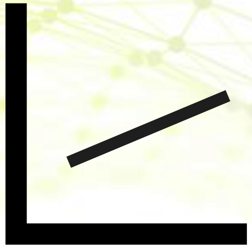
Higher values indicate more variability.

Can be used to standardised the variability of a sample to compare across datasets.

Note: p = item response propensity

Compensating for non-response bias: *Post-survey adjustments*

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION



- **Inverse probability weighting (IPW)** models the response probability via logistic regression to estimate the probability of each unit's participation.

- The inverse of these probabilities can be applied as non-response weights (Mansournia & Altman, 2016; Seaman & White, 2011).

- **Post-stratification** adjustment consists of comparing survey statistics to external benchmarks.

- Survey weights are calculated so that the weighted sample aligns with **known population distributions** across specific, **aggregate-level subgroups.**

Compensating for non-response bias: *Post-survey adjustments*

- **Raking** (or iterative proportional fitting) repeatedly adjusts sample weights so that weighted marginal distributions match relevant population totals (Deville and Särndal,1992).

- Raking requires only **marginal distributions** but can yield high within-household heterogeneity.

- **Calibration** weighting follows a similar iterative proportional fitting approach but constrains weights to be equal within households.

- The weighted distribution of household size exactly matches the number of individuals in the weighted sample (National Centre for Research Methods, 2012).

Compensating for non-response bias: *Imputation methods*

- Imputation involves replacing missing data with substituted values from auxiliary data sources, allowing for complete case analyses.

Single Imputation methods fill in missing values with a single estimate, such as:

- ***Mean/Median Imputation*: Replacing missing values with the mean or median of observed responses.
- ***Regression Imputation***: Predicting missing values using regression models based on auxiliary variables.
- ***Nearest Neighbour***: Assigning missing values based on the closest observed data point in terms of similarity on a set of relevant variables.
- ***Hot Deck Imputation***: Substituting missing values with observed responses from a random similar unit within the dataset.

Compensating for non-response bias: *Imputation methods*

Multiple Imputation (MI) creates multiple complete datasets, merging results to produce estimates that account for both within-and between-imputation variability (Rubin, 1987) for example:

- ***Multiple imputation via chained equations***: Using repeated regression modelling to draw a missing value from a random distribution and assumes each imputed variable is conditional on all other variables (Raghunathan, Lepkowski, Van Hoewyk & Solenberger, 2001).

- ***Multiple imputation via predictive-mean matching***: Instead of drawing an imputed value from a random distribution, it is possible to draw an observed value from a donor having a similar predictive mean (Morris, White & Royston. 2014).

Considerations for *compensating for non-response bias*

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- These methods assume **MAR** conditional on the chosen auxiliary variables and depend on the accuracy of integrated population margins.
- Non-response weighting often results in a trade-off with increased variance.
- Imputation and weighting models need careful specification and may be computationally intensive, especially with large datasets or numerous variables **(White et al., 2011).**
- Datasets containing imputed values should not be thought of a observed data, but as a methodology for conducting statistical analyses that adjust for non-response biases **(National Centre for Research Methods, 2012).**

What can integrated data offer?
Non-survey data

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Census data
  - UK census data provided by the Office for National Statistics (2025) is a gold standard for weighting and post-stratification of survey estimates.
  - Mid-year population estimates are updated annually to adjust for births, deaths and migration patterns, among other factors (Office for National Statistics, 2025).

- Administrative data
  - Administrative records are collected for routine and operational purposes **(Harron et al, 2017),** including:
    - **Health data (***Hospital episode statistics;* ***Rajah et al., 2023).***
    - **Education data** ***(****Educational records;* ***Booth et al., 2024).***
    - **Employment data** (*Employment spells;* ***Büttner, Sakshaug & Vicari, 2021).***

- Geospatial data
  - **Geospatial characteristics** can provide more granular detail on **contextual geographical factors**, and **stratification variables** (e.g. the **Postcode Address File; PAF** and the **Census**).
  - Integrated survey and geospatial data is particularly useful for weighting and calibration methods to spatially rebalance the survey sample (Office for National Statistics, 2022).

What can integrated data offer?
Survey-related data

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Sampling frame data
  - The **Postcode Address File (PAF)** being the most commonly used to derive sampling frames for UK surveys.
  - Sampling frame information is particularly relevant for **cross-sectional or first-wave longitudinal studies**, which cannot rely on previous wave comparisons.
- Survey paradata
  - Survey paradata describes data about the survey process **(Blom, 2008).**
  - Contact data is available for both **respondents** and **non-respondents** and is often related to both the survey process and the survey outcome **(Kreuter, Lemay and Casas-Cordero, 2007; Blom, 2008).**
- Other survey waves
  - Whether a cohort or panel member **responds at previous waves** is often among the strongest predictors of current and future non-response **(Silverwood et al., 2024).**
  - Linking **prior wave variables** with **current wave response** outcomes can serve as indicators for non-response weighting and predictors in multiple-imputation models.

Recommendations

Use the response rate alongside indicators of representativity (R-indicators) and variability (CVs) **(Groves, 2006; Maslovskya et al., 2025).**

When selecting indicators, use a generic set to and a survey specific set to compare across and within data sources **(Statistics Netherlands, 2025; Maslovskya et al., 2025)**

Start with a limited set of well-measured variables, and document each step and diagnostics to ensure FAIR inference **(National Centre for Research Methods, 2012; Wilkinson, 2016).**

Imputation and weighting are model specific and should be tailored for each analysis. The CLS missing data strategy **(Silverwood, 2024)** provides an overview of useful steps.

Summary

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

1. What is non-response bias?

When non-respondents are systematically different to respondents on key survey variables.

2. Methods for evaluating and compensating for non-response bias

Evaluating:

- *Population total comparisons, sub-group comparisons.*
- *R indicators and coefficients of variation.*

Compensating:

- *Inverse probability weighting, post-stratification, raking, calibration.*
- *Single and multiple imputation.*

3. Additional data sources for handling non-response bias

- *Census benchmarks*
- *Administrative records*
- *Geospatial data*
- *Sampling frame data*
- *Survey paradata*
- *Other survey waves*

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

RS7: Data Integration

Prof Alexandru Cernat (RS lead, Uni. Manchester)

Prof Natalie Shlomo (Uni. Manchester)

Prof Nikos Tzavidis (Uni. Southampton)

Prof Joseph Sakshaug (IAB)

Dr Thomas O'Toole (Uni. Manchester)

thomas.otoole@manchester.ac.uk

How is non-response defined?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- The measurement of non-response in survey data traditionally focusses on response-rate
- The response rate can be calculated as "the number of complete interviews with reporting units divided by the number of eligible reporting units in the sample" **(AAPOR, 2023).**
- The inverse of the response rate is the non-response rate.
- The response rate is the property of a survey, whereas non-response bias is the property of a statistic **(Wagner, 2012).**

$$RR = \frac{n_r}{n_e}$$

$$NRR = 1 - \frac{n_r}{n_e}$$

Note: n_r = complete interviews with reporting units, n^e = number of eligible reporting units in the sample frame

How is non-response bias defined?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- The response rate is the property of a survey, whereas non-response bias is the property of a statistic **(Wagner, 2012).**

- The **deterministic** non-response bias of a mean value can be calculated by the non-response rate multiplied by the difference between the estimates of the survey respondents and non-respondents **(Groves, 2006; Koch & Blom, 2016).**

$$NRB(\bar{y}) = NRR * (\bar{y}_R - \bar{y}_{NR})$$

Note: $\bar{y}$ = respondent set mean, NR = non-response, NRR = non-response rate, NRB = non-response bias.

- ***However, the difference between respondents and non-respondents on a variable of interest is often difficult to ascertain (via population statistics) or unknown.***

How is non-response bias defined?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

According to **Groves & Peytcheva (2008)**, the decision of a respondent to refuse or attrit is thought to be conditional on:

- **Separate causes** (e.g. discrete health events)
- **Common causes** (e.g. socio-demographics)
- **Survey variable causes** (e.g. length and topic)

Non-response bias can be estimated and calculated via a **stochastic** formula as the ratio between the covariance of the survey variable, and survey variable response propensity, and the mean response propensity; response rate **(Groves 2006; Koch & Blom, 2016).**

$$NRB(\bar{y}) = \frac{\sigma_{(y,p)}}{RR}$$

Note: $\bar{y}$ = respondent set mean, RR = response rate, NRB = non-response bias.

The challenge of non-linkage

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- In the context of data integration, an important mechanism of missingness is non-linkage; **non-consent and linkage error.**

- **Linkage non-consent** refers to where consenting survey participants will not opt-in to provide a common unit across data sources and are unable to be linked **(Sakshaug, 2021).**

- Non-consent differences bear closer to the respondent's level of trust in the linked data provider **(Jäckle, Burton, Couper, Crossley & Walzenbach, 2021).**

- **Linkage error** refers to missed and incorrect matches between survey and non-survey units.

- When the matching procedure is deterministic missed matches are more common.

- If the matching procedure is probabilistic, incorrect matches are more likely **(Harron et al., 2017).**