



# **SURVEY FUTURES**

**SURVEY DATA COLLECTION  
METHODS COLLABORATION**

## **Report 15: How could the Reference Data Management Framework improve survey quality?**

**Zac Marco Perera, Gerry Nicolaas & Nick Waugh**

National Centre for Social Research  
with support from the Office for National Statistics

June 2026

**[www.surveyfutures.net](http://www.surveyfutures.net)**

Survey Futures is an Economic and Social Research Council (ESRC)-funded initiative (grant ES/X014150/1) aimed at bringing about a step change in survey research to ensure that high quality social survey research can continue in the UK. The initiative brings together social survey researchers, methodologists, commissioners, and other stakeholders from across academia, government, private and not-for-profit sectors. Activities include an extensive programme of research, a training and capacity-building (TCB) stream, and dissemination and promotion of good practice. The research programme aims to assess the quality implications of the most important design choices relevant to future UK surveys, with a focus on inclusivity and representativeness, while the TCB stream aims to provide understanding of capacity and skills needs in the survey sector (both interviewers and research professionals), to identify promising ways to improve both, and to take steps towards making those improvements. Survey Futures is directed by Professor Peter Lynn, University of Essex, and is a collaboration of twelve organisations, benefiting from additional support from the Office for National Statistics and the ESRC National Centre for Research Methods. Further information can be found at [www.surveyfutures.net](http://www.surveyfutures.net).

Research Strand 1 of Survey Futures (“Enhanced sampling frames and procedures”) focuses on the challenges associated with our dependency on using the Postcode Address File as a sampling frame for random probability surveys. One of the projects within this strand, commissioned under Phase 2 of Survey Futures and led by Gerry Nicolaas, NatCen, explores the feasibility of using the Reference Data Management Framework produced by the Office for National Statistics to improve survey sampling processes, as well as other design features and survey implementation.

The suggested form of citation for this report is:

Perera Z M, Nicolaas G & Waugh N (2026) ‘How could the Reference Data Management Framework improve survey quality?’, *Survey Futures Report* no. 15. Colchester, UK: University of Essex. Available at <https://surveyfutures.net/reports/>.

## Contents

1	Introduction.....	6
2	Objective and research questions .....	7
3	Reference Data Management Framework .....	8
4	Methods.....	10
4.1	Statistical quality.....	10
4.2	Non-statistical quality.....	12
5	Impact on representation .....	13
5.1	Defining target populations to minimise specification error .....	13
5.2	Constructing sampling frames to minimise coverage error.....	14
5.3	Designing samples to minimise sampling error .....	16
5.4	Amending fieldwork design to mitigate non-response error.....	17
5.5	Computing weights to adjust for coverage, sampling, and non-response errors	19
6	Impact on measurement .....	22
6.1	Operationalising constructs to maximise measurement validity.....	22
6.2	Designing questionnaires to minimise measurement error .....	23
6.3	Pre-testing items to minimise measurement error.....	24
6.4	Processing data to minimise measurement error .....	25
7	Impact on non-statistical quality.....	27
7.1	Relevance .....	27
7.2	Timeliness.....	28
7.3	Credibility .....	29
7.4	Completeness.....	30
7.5	Coherence and comparability.....	30
7.6	Accessibility and usability .....	32
8	Constraints and possible solutions in using the RDMF to improve survey quality..	33
8.1	Inclusivity and representativeness .....	34
8.2	Quality and transparency of linkage .....	35
8.3	Conceptual and definitional misalignment .....	36
8.4	Quality of underlying administrative sources .....	37
8.5	Governance, permissions, and legal constraints .....	38
8.6	Resourcing, skills and capacity constraints .....	39
9	Conclusion.....	40
	References .....	43

Appendix A: International use of population registers in social surveys ..... 47  
Appendix B: Public acceptability of administrative data use in research and implications  
for the RDMF..... 52

## **Acknowledgement**

We would like to express our sincere gratitude to the Office for National Statistics (ONS) for their invaluable support, collaboration, and contributions to the preparation of this report. Their collective expertise, insights, and assistance have significantly enhanced the quality and comprehensiveness of this work. In particular, we acknowledge their substantial contribution through the addition of chapter 8 and appendices A and B, which have enriched the report considerably. We are deeply appreciative of their efforts and dedication.

## 1 Introduction

High-quality social survey data are essential for the production of robust statistics that inform policymaking and social research. However, social surveys in the United Kingdom (UK) are increasingly challenged by declining response rates, rising data collection costs, and growing concerns over data quality and representativeness (Office for National Statistics, 2025).

Developments in the field of survey research offer promising solutions to these challenges, such as advanced techniques for targeting and tailoring survey designs to specific populations, greater use of online data collection, and the integration of administrative data to enhance data accuracy and depth (Couper, 2017). However, the UK faces restrictions in fully capitalising on these new advancements because it lacks a population register – an official database that records basic information about all residents. Instead, reliance on the Postcode Address File (PAF) for address-based sampling poses limitations as it only provides address information. As a result, specific populations cannot be identified for targeted or tailored approaches, email contact cannot be used for online data collection, and administrative data cannot be linked for all sampled cases including non-respondents to assist in weighting for non-response. Moreover, PAF excludes people living outside private residential addresses.

Although it is unlikely that the UK will have a population register for the foreseeable future, the Office for National Statistics (ONS) is developing the Reference Data Management Framework (RDMF), which is built from administrative data sources (Office for National Statistics, 2021). While the RDMF is primarily intended for data linkage, it may also offer opportunities to support aspects of survey design and potentially improve data quality. This paper considers how the RDMF might reduce the need for within-household sampling, thereby reducing clustering effects and possibly enabling more effective oversampling of sub-groups. Additionally, the RDMF could facilitate the targeting and follow-up of underrepresented groups with tailored fieldwork efforts, which may reduce the risk of non-response bias. Linking the RDMF with other administrative data may allow some questionnaire content to be shortened, potentially lowering respondent burden and improving data quality, for example, in the measurement of income. Finally, the RDMF may be able to provide more relevant covariates for imputation and non-response weighting, which could improve the accuracy and credibility of survey estimates.

If the RDMF could be used, it could potentially simplify survey data collection and produce higher quality statistics that better meet data user needs, more quickly and at a lower cost. This potential is amplified by the opportunities presented by data integration with administrative records, including the production of more granular and inclusive statistics. However, the feasibility of using the RDMF for survey design and implementation could be constrained by data quality issues, legal and ethical considerations, technical challenges, and practical implementation. It should be noted that the RDMF is an ONS product, and any use

of it by other organisations is most likely through indexed datasets rather than through direct access to the RDMF.

## 2 Objective and research questions

The objective of this study is to theoretically explore the potential and limitations of using the Reference Data Management Framework (RDMF) to enhance the design and implementation of social surveys in the UK. This exploration is guided by two primary research questions:

- (1) How could the RDMF be used to improve the end-to-end survey data collection process?
- (2) What are the constraints and possible solutions for using the RDMF in the design and implementation of social surveys?

This report lays the groundwork for a more detailed assessment of the theoretical quality improvements and cost reductions achievable using the RDMF. This will be achieved by simulating a redesign and re-costing of the Wales National Travel Survey<sup>1</sup>, a mixed-mode survey that currently uses the Postcode Address File (PAF) as a sampling frame. This approach will enable a direct cost-quality comparison between using the RDMF and the PAF in survey design and implementation. This work will be published in a separate *Survey Futures* report.

---

<sup>1</sup> The Wales National Travel Survey is a cross-sectional survey which collects data on travel attitudes and behaviour from a random probability sample of people living in Wales. The survey is carried out by the National Centre for Social Survey on behalf of Transport for Wales.

### 3 Reference Data Management Framework

The Reference Data Management Framework (RDMF) is a centralised infrastructure that allows the ONS to link its data with administrative data from other government departments, ensuring secure, consistent, and efficient management of reference data across its various statistical and analytical processes, enabling more comprehensive and cohesive analysis (Office for National Statistics, 2026).

The RDMF is a collection of datasets that have been linked together to create four indexes, consisting of three core indexes called Demographic, Business, Location, and one supporting index called Classification:

1. **The Demographic Index (DI)** is a longitudinal dataset from 2016 onwards, created by linking data on citizens' interactions with public services. The DI brings together administrative datasets from a range of government departments and authorities to ensure good coverage of the population, such as birth and death registrations, HMRC Frameworks, the Personal Demographics Service, Welsh School Census, English School Census, Lifelong Learner Wales Record, Individualised Learner Record, Higher Education Statistics Agency. The population coverage for Scotland and Northern Ireland is more limited than for England and Wales due to the absence of equivalent health and education data.
2. **The Business Index (BI)** provides information for businesses and organisations (at Legal Unit levels) across the UK, including foreign businesses registered in the UK for trade purposes. It includes every variable from HMRC VAT & PAYE, plus data from the Financial Conduct Authority and Companies House. There are plans to include other data sources from HMRC, Companies House, the Charities Commission, and the Global Legal Entity Identifier.
3. **The Location Index (LI)** is a composite of the Address and Geography Indexes.
  - The Address Index (AI) contains all the addresses in the UK (with unique property reference numbers). It includes data from the AddressBase Premium and Islands, the Postcode Address File, the National Address Gazetteer, and the quarterly NRS Address Data dataset from National Records of Scotland.
  - The Geography Index (GI) contains a range of statistical and administrative geographies created by the ONS Geography team, such as Output Areas, Local Authority Districts, Ward Boundaries, Parliamentary Constituencies, Regions, Health Geographies, and Postcode Boundaries.
4. **The Classification Index (CI)** is called a supporting index because it includes industry and occupation lookups to support the DI and the BI. It includes classifications such as

Standard Industrial Classification (SIC) and Standard Occupational Classification (SOC) that are used throughout the Government Statistical Service.

The three core indexes contain one entry for each person, organisation, and location in the country. Identifying information such as names, addresses, and national ID numbers have been removed and replaced with a reference number, to ensure that people cannot be directly identified. These reference numbers are used to link data together via matching.

The RDMF is still in development with the various indexes and services at different stages of maturity. Work at ONS is ongoing to learn more about the quality of the indexes, but the following benefits have already been realised:

- Enhanced security, standardisation, and ease of access for researchers.
- Robust development processes, including thorough unit testing and methodological review.
- Almost complete population coverage for England and Wales through the utilisation of several large datasets. The population coverage for Scotland and Northern Ireland is more limited due to the absence of equivalent health and education data.
- The datasets chosen for the RDMF are compatible with each other using similar definitions and shared identifiers. This compatibility creates coherent data across the four indexes and improves comparability across sources.
- Timeliness, with the Business Index updated daily, the Location Index updated every six weeks, and the Demographic Index updated every six to twelve months to meet current requirements for producing population and migration statistics with plans for more frequent updates in the future (Office for National Statistics, 2026).

ONS is now focusing on measuring accuracy, especially in the Demographic Index but also the other indexes. Integrated data of this type are a novel architecture for ONS which requires new approaches to understand their quality and how the potential for error flows through the various processes to affect outputs. In addition to traditional measures of accuracy such as precision and recall, new measures need to be developed that will take ONS beyond just measuring error in the constituent data, towards a new overall methodology which will allow users to measure and communicate the quality of their analysis.

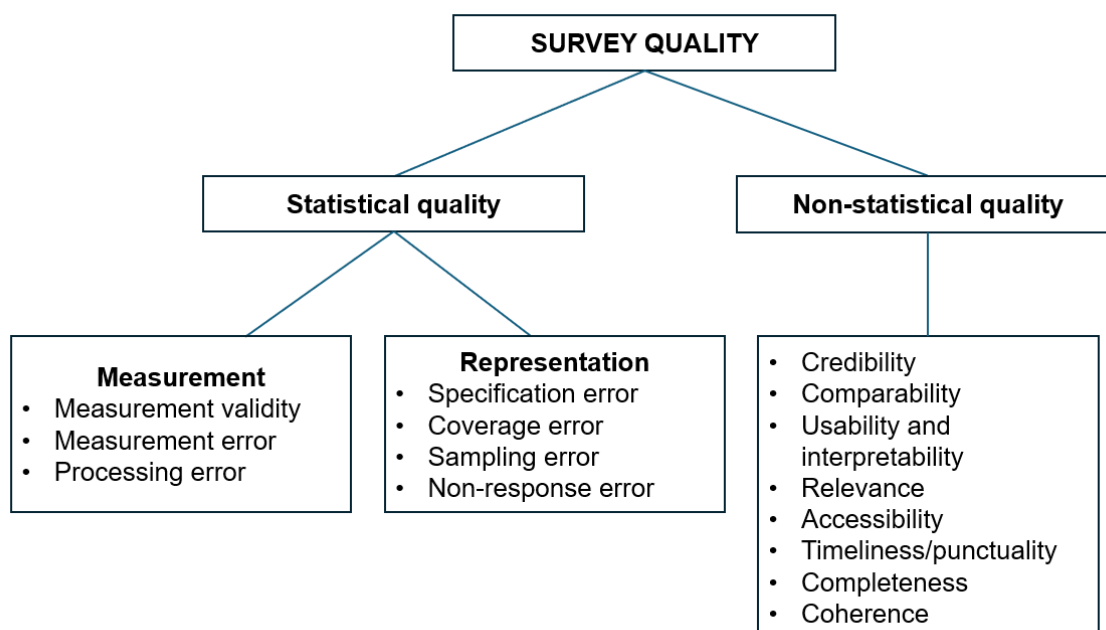
## 4 Methods

To explore how the RDMF could enhance survey quality, we have adopted a Total Survey Quality (TSQ) approach, which outlines the various potential sources of error that can occur during the survey lifecycle. These error sources compromise statistical quality as well as other dimensions of quality that relate to the needs of data users, such as credibility, comparability, usability/interpretability, relevance, accessibility, timeliness/punctuality, completeness, and coherence (Biemer, 2010).

The TSQ approach highlights the complexity of achieving high-quality survey data. Not all errors can be eliminated, and reducing one type of error may increase another type. By recognising the trade-offs involved in minimising different types of errors, we can achieve a clearer understanding of the overall impact the RDMF may have on survey statistics.

Figure 1 illustrates the different dimensions of survey quality that need to be considered when making these trade-offs.

**Figure 1.** Dimensions of survey quality.



### 4.1 Statistical quality

Survey data are considered to have high statistical quality when they produce estimates of population characteristics that are both accurate and precise. Accuracy refers to how close the survey estimates are to the true values in the target population. Precision, on the other

hand, refers to the consistency and reliability of the survey results. High accuracy means that the expected survey statistic is close to the population parameter, whilst high precision means that if the survey was run multiple times using identical methods, the results would not vary very much.

The Total Survey Error (TSE) framework offers a systematic approach for detecting the various sources of error that can compromise accuracy and precision (Groves & Lyberg, 2010). It provides a map of where errors can arise, highlights the trade-offs between different errors, and signals the ways in which quality can be improved at every stage of the lifecycle.

In the TSE framework, the error sources are grouped into two broad categories – measurement and representation. Representation error arises when the people included in the survey do not accurately reflect the target population in ways that matter to what is being measured. If representation is weak, it is not possible to confidently make generalisations about the entire population, based on the survey results. Measurement error encompasses inaccuracies that occur during the process of obtaining and handling data from respondents. This includes errors that occur when the survey questions do not accurately capture the concept intended to be measured, errors that arise when respondents misunderstand or misinterpret questions, and errors that occur during the data handling stages, such as data entry, coding, and editing. Both representation and measurement errors need to be carefully balanced and optimised, considering various trade-offs, to ensure confidence in the inferences drawn from the survey data.

When surveys are designed and implemented using the RDMF, linkage activities introduce an additional source of error that may cut across several components of the Total Survey Error framework rather than constituting a single, isolated error type. In particular, linkage error may contribute to representation error where linked administrative data are used to construct or augment the sampling frame, for example through missed or incorrect links that lead to units being omitted, duplicated, or misclassified. It is important to distinguish here between precision in the TSE framework and precision in data linkage, which refers to the proportion of true matches amongst all links.

Linkage error may also arise as a form of measurement error when survey responses are linked post-collection to administrative or other data sources, potentially distorting variables through false matches, missed links, or inconsistencies across data sources. Recognising how linkage error interacts with multiple TSE components is important when applying the RDMF, as design decisions about data integration can shift the balance of errors across the survey lifecycle and affect both statistical accuracy and broader dimensions of quality.

## 4.2 Non-statistical quality

Total Survey Quality goes beyond the statistical quality of survey data. It also depends on meeting data users' needs, such as credibility, comparability, usability/interpretability, relevance, accessibility, timeliness/punctuality, completeness, coherence (Biemer, 2010).

To effectively assess these non-statistical dimensions of survey quality, various frameworks have been developed, including those produced by the European Statistical System (ESS, 2019), the Organisation for Economic Co-operation and Development (OECD, 2011), and the Office for Statistics Regulation (Office for Statistics Regulation, 2025). Each of these frameworks categorise these different dimensions in diverse ways. Having reviewed these frameworks, we have adopted the following categories to better align with the context of our review:

- Relevance.
- Timeliness.
- Credibility.
- Completeness.
- Coherence and Comparability.
- Accessibility and Usability.

## 5 Impact on representation

This section explores how the use of the RDMF could significantly improve the representation quality of survey data. It addresses the challenges and potential solutions for defining target populations, constructing sampling frames, designing samples, improving fieldwork design, and enhancing statistical weights.

Each section delves into specific errors that affect the representation of survey data, as outlined in the Total Survey Error framework (see section 4.1):

- Defining **target populations** that better reflect the inferential population, reducing **specification error** (see section 5.1).
- Producing **sampling frames** that better reflect the target population, reducing **coverage error** (see section 5.2).
- Drawing **issued samples** that better reflect the sampling frame, reducing **sampling error** (see section 5.3).
- Improving the effectiveness of **fieldwork design** in promoting response, to mitigate **non-response error** (see section 5.4).
- Enhancing the performance of **statistical weights**, to better adjust for **coverage, sampling, and non-response errors** (see section 5.5).

By leveraging the rich and timely data provided by the RDMF, researchers could develop more accurate sampling frames, employ tailored fieldwork strategies, and use advanced weighting techniques to ensure that survey results are more representative of the target population.

### 5.1 Defining target populations to minimise specification error

Survey data are used to make inferences about a defined population of interest. Specification error occurs when there is a misalignment between the inferential population to which users wish to generalise the survey results and the target population that is formally defined at the design stage. Such errors occur at a conceptual level, before any sampling or data collection takes place, and cannot be corrected through improvements in sampling frames or estimation alone.

A common example in UK social surveys is the implicit treatment of the private household population as equivalent to the UK general population. In the absence of a national population register, many surveys define their target population operationally as people living in private households, reflecting the constraints of available sampling frames such as the small user Postcode Address File (PAF). While this definition may be pragmatic, it represents a specification error when the survey's analytical intent is to describe or compare outcomes for

the wider UK general population, including those living in communal or institutional settings. In these circumstances, exclusion is embedded in the survey definition itself, and survey findings are often inappropriately generalised by data users and policymakers without explicit recognition of, or adjustment for, systematic differences between the defined target population and the broader inferential population (Lynn, 2014).

The RDMF could offer opportunities to reduce this form of specification error by enabling target populations to be defined more closely in line with analytical objectives. In particular, administrative data accessed via the RDMF would make it feasible to conceptualise the target population as all usual UK residents, regardless of residential setting. At a minimum, this would involve drawing on the Location Index, which enables the explicit identification of both private and communal residential settings, including prisons, care homes, student halls, boarding schools, and hostels. Further reductions in specification error could be achieved by sampling directly from the Demographic Index, but this is not feasible under current access and governance constraints (see chapter 8).

By allowing survey designers to define inclusion in substantive rather than purely operational terms, the RDMF could help ensure that exclusions are a matter of transparent design choice rather than unintended specification error.

## **5.2 Constructing sampling frames to minimise coverage error**

Coverage error occurs when the sampling frame does not accurately represent the defined target population, through the underrepresentation (under-coverage) or overrepresentation (over-coverage) of eligible units. Unlike specification error, coverage error assumes that the target population has already been correctly defined and focuses instead on errors introduced during frame construction and maintenance.

For many UK surveys of the private household population, the small-user PAF is used as the sampling frame. Although it provides near-complete coverage of private households, empirical evidence suggests residual under-coverage, particularly for newly constructed dwellings, estimated at around 3% in 1991 (Foster, 1994). The frame also contains elements of over-coverage, including vacant or derelict properties, second homes, and non-residential addresses such as business premises, estimated at around 8%-10% (NatCen, unpublished data, 2026). These frame imperfections affect the probability that eligible units are included and sampled, but do not in themselves imply any mis specification of the target population.

Using the Location Index from the RDMF for sampling would represent a marked improvement over PAF in terms of both under-coverage and over-coverage. By integrating information from The National Address Gazetteer and the Ordnance Survey datasets, the Location Index

reduces the likelihood of missing new builds and properties without mail delivery and allows better identification of demolished and retired addresses. Conditional on a target population that includes all eligible residential settings, the Location Index supports the identification and inclusion of communal and institutional residences alongside private addresses. Over-coverage can be further reduced by linking the Location Index to the Business Index to identify and remove addresses associated solely with business activity from the sampling frame, while retaining institutional residential addresses where people live.

In principle, further reductions in coverage error could be achieved if it were possible to sample individuals directly from the Demographic Index rather than indirectly through addresses in the Location Index, thereby shifting the sampling frame from addresses to individuals. While an address-based frame improves coverage of residential settings, it does not fully resolve person level coverage problems, such as individuals who are weakly or inconsistently associated with addresses. An individual level sampling frame would enable eligibility and selection probabilities to be defined directly for people rather than inferred from address characteristics, thereby reducing under-coverage of hard-to-locate individuals. The extent to which these coverage gains could be realised would depend on data access, governance constraints, and the quality of linkage underpinning the Demographic Index. In practice, however, implementation would still depend on address-based contact, as the RDMF does not include direct contact details such as email addresses or phone numbers. An individual-level sampling frame would therefore still rely on addresses as the means of contacting selected individuals.

A related form of coverage error arises when a unit appears more than once in the sampling frame (over-coverage). When sampling addresses from either the PAF or the Location Index, individuals with more than one residence have a higher probability of selection. If it were possible to sample individuals directly from the Demographic Index or to link sampled addresses from the Location Index to the Demographic Index, it would become possible to identify individuals associated with multiple addresses and to account for this through design choices, eligibility rules, or weighting adjustments, thereby mitigating the impact of such over-coverage.

Finally, the RDMF itself may exhibit some degree of coverage error due to temporal lags, such as delays in removing records for individuals who have died or emigrated and delays in records being updated when circumstances change. The extent of such errors will depend on update frequencies and linkage processes across administrative sources. These limitations highlight that, while the RDMF has the potential to improve coverage relative to traditional frames, careful evaluation of frame quality remains essential even when the target population is correctly specified.

### 5.3 Designing samples to minimise sampling error

Sampling error is the difference between an estimate derived from a sample survey and the value that would have been obtained if a census had been conducted among the entire population. The key drivers of sampling error are sample size and the sample design.

Sampling error can be reduced by sufficiently increasing the sample size to capture the diversity and variability within the population. This could be further enhanced by dividing the sampling frame into strata based on specific characteristics before drawing samples from each stratum, thereby ensuring that each subgroup is adequately represented in the final sample. Additionally, oversampling can be employed within one or more of these strata to ensure that smaller sub-groups, which may not be adequately represented in a proportionate sample, have enough cases for meaningful analysis.

In the UK, address-based sampling using the Postcode Address File (PAF) is commonly used, and the information available for stratification is largely limited to area-level data, such as region, Index of Multiple Deprivation (IMD), and census-based variables that provide proportions with specific characteristics in an area such as the percentage of households with a non-manual head of household. While effective to some extent, such proxies are only imperfectly correlated with individual level characteristics and survey outcomes.

As the RDMF is a framework designed to facilitate the linkage of administrative data held in other sources, it could, in theory, enable access to a wider range of records containing individual-level characteristics such as age, sex, income, ethnicity, health, and education. If permitted, and subject to appropriate legal, ethical and governance considerations, the use of such information for stratification purposes could reduce sampling error, particularly when the selected characteristics are strongly correlated with survey outcomes.

In addition to sample size and design, sampling error in practice can also be affected by how selection procedures are implemented. A limitation of address-based sampling in the UK is that it does not permit the direct selection of individuals, which is a constraint for individual-based surveys but less so for household surveys. Once contact has been made at a sampled address - either in person or in writing - a separate procedure is required to randomly select eligible respondents. For interviewer-administered individual-based surveys, interviewers are given instructions to ensure that a random household is selected in the case of multi-household addresses, and subsequently, one or more randomly selected adults within the household. For self-administered surveys, instructions are provided in the letter for the householder to follow. Commonly used techniques include the Kish grid, CAPI randomisation, the next/last birthday method, and up-to-two/three/four adults (Smyth, Olson & Stange, 2019).

In practice, these within-household selection methods are vulnerable to arising from human error or non-compliance. This risk is particularly pronounced for self-administered surveys (Reece & Lynn, 2015), which are becoming an increasingly common mode of data collection over time (Cabrera- Álvarez, et al., 2025).

Subject to appropriate legal, ethical and governance considerations, the RDMF could support the selection of individuals rather than addresses. This could be achieved by sampling individuals directly from the Demographic Index or by linking individual-level entries from the Demographic Index to addresses drawn from the Location Index. Either approach would allow individuals to be selected explicitly at the design stage, removing the need for within household respondent selection during fieldwork and reducing the additional sampling variability and error associated with these procedures.

#### **5.4 Amending fieldwork design to mitigate non-response error**

In any survey, it is crucial that the achieved sample—the group of respondents who participate—closely resembles the issued sample, which is the group initially selected to take part. Non-response error arises when non-respondents differ systematically from respondents in ways that are relevant to the survey outcomes, potentially distorting the survey findings and leading to invalid conclusions.

There are two principal ways in which the RDMF could contribute to mitigating non-response error: first, by supporting the oversampling of groups that are known to be less likely to participate, and second, by enabling more tailored and adaptive fieldwork strategies aimed at improving response rates and reducing differential non-response.

##### *Oversampling hard-to-survey groups*

Firstly, the RDMF could be leveraged to oversample sub-groups that are known to be less likely to participate in surveys. By drawing on linked administrative data, researchers could in principle identify characteristics such as age, income, health status, or past interactions with public services that are associated with non-response and increase the selection probabilities of these groups within the issued sample. This approach may help ensure that hard-to-survey groups are more adequately represented in the achieved sample.

However, the effectiveness of this approach depends on the quality and completeness of linkage within the RDMF. Some groups may be more difficult to link accurately and may therefore be underrepresented or misclassified in the administrative data, limiting the scope for effective oversampling.

In addition, oversampling alone does not eliminate non-response bias if respondents differ systematically from non-respondents in other ways that affect key survey variables.

### *Tailoring fieldwork efforts*

Secondly, the RDMF could reduce non-response error by supporting more tailored fieldwork strategies aimed at improving response rates and reducing differential non-response. Fieldwork efforts include survey communications, modes of data collection, incentive strategies, and adaptive and responsive survey designs, each of which could potentially be informed by linked socio-demographic information available through the RDMF.

The RDMF could support more effective *survey communications* by enabling sampled cases to be linked to socio-demographic characteristics derived from other approved data sources. This information could be used to tailor the content, tone, or frequency of communications for different groups, for example by emphasising particular motivations for participation or adjusting reminder strategies for groups with historically lower response rates. Such tailoring could improve the relevance and salience of communications.

The effectiveness of these approaches would likely be enhanced further if access to individual names were possible, as personalised communications are consistently associated with higher response rates than generic invitations, as they strengthen legitimacy, personal relevance and individual responsibility to respond (Dillman, Smyth & Christian, 2014). However, using the RDMF to obtain names would require substantial changes to governance arrangements and data-sharing agreements (see section 8).

Socio-demographic information linked via the RDMF could also be used to assign *different modes of data collection and different mixes* of these modes to groups based on the format they are most likely to participate in. For example, groups known to have lower levels of digital access or literacy (e.g., those aged 80 and over) could be offered interviewer-administered modes rather than web-only options. While mixed-mode designs introduce a risk of measurement differences, careful questionnaire design and harmonisation can mitigate these effects, and evidence suggests that the benefits for representativeness often outweigh the associated disadvantages (Maslovskaya, et al., 2025).

The RDMF could also be used to tailor *incentive strategies* by enabling differential incentives to be targeted at groups with persistently lower response rates. For example, higher-value incentives could be offered to those in receipt of government benefits. Differential incentives have the potential to reduce non-response bias by increasing participation among under-represented groups, thereby improving the balance of the achieved sample.

Finally, the RDMF could play a role in both *adaptive and responsive survey design*. In adaptive designs, information linked via the RDMF could be used prior to fieldwork to identify characteristics associated with non-response and to tailor survey protocols accordingly. In responsive designs, response rates could be monitored across linked characteristics during data collection, allowing resources to be dynamically reallocated—for example through increased follow-up or alternative contact strategies—if particular groups are found to be under-represented. This ability to anticipate and respond to emerging patterns of non-response has the potential to further enhance survey quality and reduce non-response error.

## **5.5 Computing weights to adjust for coverage, sampling, and non-response errors**

Statistical weighting is the final phase of survey design in which the survey data are adjusted to better represent the target population. Weights adjust for compositional differences between the achieved sample and the population of interest, arising from coverage, sampling, and non-response errors.

*Design* weights correct for unequal selection probabilities arising from the sample design (for example, due to oversampling of specific groups), while *calibration* or *post-stratification* weights are used to adjust the responding sample to align with known population distributions, thereby mitigating the effects of coverage and non-response errors. While applying weights will improve accuracy, it will also reduce precision by increasing uncertainty in the estimates. The extent of this loss of precision depends on both the magnitude of the representation errors and the degree to which the resulting weights vary.

One way to improve the quality of survey estimates is to reduce the extent to which the weights must compensate for representation errors. This can lead to better precision in the survey estimates because the weights do not need to be as extreme, thereby reducing variability in the weighted data. For example, by constructing more accurate sampling frames with the RDMF, coverage errors are reduced, which decreases the variability in weights and lowers the design effect, thereby increasing the effective sample size. Similarly, by sampling individuals directly rather than households for individual-based surveys, design weights do not have to compensate for unequal probabilities of selection due to household size variation, also increasing the effective sample size. Finally, reducing non-response errors through targeted and tailored interventions ensures that the sample more accurately reflects the population, further enhancing the effectiveness of post-stratification weights.

In addition to reducing the need for large weighting adjustments, the RDMF could improve the effectiveness of the weighting process in increasing the accuracy of survey estimates. Currently, the weighting of survey data in the UK relies on area-level data from the decennial

census, mid-year population estimates and large-scale household surveys such as the Labour Force Survey and the Annual Population Survey. The RDMF could provide access to a wider range of relevant covariates at the individual level that are more timely, accurate, and granular with near-complete coverage (see section 3).

- **Timeliness of Data:** Although the RDMF's Demographic Index is currently updated every six to twelve months, plans for more frequent updates in the future will provide access to more timely data for weighting. This is advantageous because it ensures that the weighting process is based on the most current population information, improving the relevance and accuracy of the survey estimates and better reflecting recent demographic changes.
- **Range of Variables:** Using the RDMF for survey weighting could provide access to a broader set of covariates derived from administrative data. This would allow weighting models to incorporate variables that are more closely associated with both survey participation and the outcomes of interest, rather than relying on a limited and standardised set of variables typically available when using PAF samples. This reduces the risk of missing important variables in the weighting models, thereby enhancing the effectiveness of the weights in improving the accuracy of substantive indicators.
- **Consistency of Covariates:** Weighting requires information on survey respondents that is also available for the population as a whole. Typically, this information is collected in the survey. If administrative data accessed via the RDMF could be used instead, this would ensure that the same covariates are defined and measured consistently for respondents and nonrespondents alike. This consistency reduces discrepancies between survey and population benchmarks, improving the effectiveness of weighting and reducing the risk of bias.
- **Granularity of Data:** Having access to administrative records would offer more granular data compared to aggregated data from traditional sources such as the Census and ONS mid-year population estimates. This includes finer geographic resolution and more detailed categorisations of characteristics such as income, employment, health, or education. Such granularity allows for more precise adjustments in the weighting process, leading to more accurate and representative survey estimates.
- **Individual-Level Data:** Having access to individual-level data through the RDMF would be a significant advantage for weighting compared to area-level data from traditional sources. Individual-level data allows for more precise adjustments, as it directly reflects the characteristics of each person rather than aggregated averages. This precision enhances the accuracy and representativeness of the survey estimates, as it

better accounts for variability within the population and reduces the risk of biases that can arise from using broader area-level data.

In summary, reducing the extent to which weights must compensate for representation errors is crucial for improving the precision and reliability of survey estimates. The RDMF offers significant advantages to improve the effectiveness of weighting by providing access to much richer and more timely data at the individual-level, rather than relying on area-level data, and a wider range of variables that could potentially be better correlated with both survey participation and key survey variables of interest.

## 6 Impact on measurement

This chapter explores the ways in which the RDMF could enhance the measurement quality in surveys. It delves into the key issues related to operationalising constructs, designing questionnaires, pre-testing items, and processing data to maximise measurement validity and minimise measurement error. The RDMF offers opportunities to incorporate more accurate and comprehensive indicators, reduce respondent burden, and improve the precision and reliability of survey estimates. Each section addresses specific strategies and benefits of utilising the RDMF to enhance the overall measurement process in surveys.

At each stage of the survey process, the RDMF could improve measurement quality by:

- Incorporating a wider set of indicators to measure underlying constructs during **operationalisation**, increasing **measurement validity** of final estimates (see section 6.1).
- **Designing questionnaires** that are shorter and include less demographic and socio-economic items, reducing **measurement error** (see section (6.2).
- Strengthening questionnaire refinement during **pre-testing**, further reducing **measurement error** (see section 6.3).
- Improving the effectiveness of **editing, validation, and imputation**, reducing **processing error** (see section 6.4).

### 6.1 Operationalising constructs to maximise measurement validity

Operationalisation is the process whereby underlying constructs of interest are defined and translated into one or more measurable indicators. Measurement validity depends on how well these indicators capture the intended construct, independent of the specific instruments used to collect them. The RDMF could potentially support improved operationalisation by enabling a broader and potentially more accurate set of indicators to be combined when measuring complex social and economic concepts.

For example, the construct of a household's income security could be operationalised using multiple indicators, such as household income over the last five years and the ability to put money aside for unexpected expenses. Rather than relying exclusively on self-reported survey measures for all components, the RDMF could enable administrative payroll data to be used as an indicator of longer-term income patterns, while survey questions could be used to capture subjective or experiential aspects that are not observable in administrative sources, such as perceived financial resilience. An example of the latter is the question "Are you able to put money aside to cover unexpected expenses?", as used in the Family Resources Survey (Department for Work & Pensions, 2025).

Under this approach, administrative and survey indicators could be combined not to replace survey questions per se, but to strengthen the conceptual measurement of the construct itself. Administrative payroll data offer advantages in terms of accuracy and completeness compared with self-reported income data, which are prone to item nonresponse and recall error, while survey questions remain essential for measuring perceptions, attitudes, and experiences. By enabling the integration of complementary indicators drawn from different data sources, the RDMF could improve measurement validity relative to relying solely on survey-based indicators.

## **6.2 Designing questionnaires to minimise measurement error**

Once constructs have been operationalised into survey variables, they are transformed into questionnaire items. Measurement error arises at this stage when the recorded response is different to the true value. This could be due to poorly designed questions and response options, differences in responses caused by the survey mode, and respondent behaviour. Questionnaire design plays a critical role in limiting this type of error.

Established principles of questionnaire design emphasise clarity, neutrality, and simplicity, ensuring that questions are interpreted consistently across respondents and that response options are balanced and exhaustive (Bradburn, Sudman, & Wansink, 2004). In multi-mode surveys, questions should also be designed to minimise mode-specific measurement error arising from interviewer effects, visual presentation differences, or satisficing behaviour (d'Ardenne, Bull, Das, Perera, & Sexton, 2025).

The RDMF could potentially contribute to reducing measurement error at the questionnaire stage by enabling certain survey questions that are prone to respondent error to be replaced with equivalent data from administrative records. While government surveys may rely on “public task” as the legal basis for such linkage under GDPR legislation, non-government surveys would typically need to obtain respondent consent. Through linkage to approved administrative records, variables such as sex, age, income, and use of public services could be obtained without requiring direct respondent reporting. For example, income is frequently under-reported or misreported in surveys by respondents at both ends of the earnings distribution in surveys (Valet, Adriaans, & Liebig, 2019). Replacing or supplementing survey income questions with administrative data from HMRC could therefore reduce measurement error attributable to misreporting. However, this may involve a trade-off with timeliness, as tax records are often only available some time after the reference period. It may also raise issues of completeness and consistency because coverage and definitions can differ across administrative systems.

Secondly, replacing survey questions with data from administrative records could reduce the overall length of the questionnaire and respondent burden, which in turn may lower measurement error elsewhere in the survey. When respondents perceive a survey to be overly long or complex, they may experience fatigue, frustration, or disinterest, leading to a decline in the accuracy and completeness of their responses. This could result in skipped questions, survey satisficing (Krosnick, 1991), or providing socially desirable answers rather than true reflections of their opinions or behaviours.

This approach was successfully implemented in the 2021 Census, where the item asking for the number of rooms was replaced with administrative data from the Valuation Office Agency (Office for National Statistics, 2020). Similarly, in the National Travel Survey, detailed questions on vehicle registration are skipped when respondents consent to obtain this information from the Driver and Vehicle Licensing Agency (DVLA) database (Department for Transport, 2025).

### **6.3 Pre-testing items to minimise measurement error**

Pre-testing of questionnaire items often involves qualitative methods, such as cognitive interviewing. Participants are typically selected based on specific characteristics that align with the target population or key subgroups within it, using purposive sampling rather than random probability sampling to ensure diversity and representation of different perspectives and understanding. This approach helps to identify question design issues that may affect subgroups differently, including problems of comprehension, recall, judgement, or response mapping.

The RDMF could support the recruitment of participants for pre-testing by enabling the identification of individuals with characteristics relevant to the cognitive testing objectives, using linked administrative records. This could improve the efficiency and precision of participant selection.

In addition, administrative data accessed through the RDMF could be used as a reference point when evaluating responses provided during pre-testing, rather than being treated as definitive or error-free measures. Administrative records are themselves subject to error, incompleteness, definitional differences, and temporal lags, and therefore should not be assumed to represent a “true” value. However, apparent discrepancies between survey responses and linked administrative data can be analytically valuable, prompting targeted probing during cognitive interviews to explore how respondents interpret questions, retrieve information, and formulate answers.

These examples illustrate how administrative data, when used cautiously and reflexively, can enhance pre-testing and reduce questionnaire-induced measurement error, while recognising that both survey responses and administrative records are imperfect representations of underlying constructs.

#### **6.4 Processing data to minimise measurement error**

Once raw survey data have been collected, they are processed to ensure that the data are internally consistent, coherent, complete, and ready for use. Data processing typically involves identifying duplicate or potentially fraudulent records, detecting missing values, verifying responses against known valid ranges, detecting outliers, resolving logical inconsistencies, and coding open-ended responses. This is typically achieved by examining the raw survey data and applying predefined rules. If errors remain undetected or the predefined rules are not applied correctly, measurement error may be introduced or amplified during processing.

The RDMF could enable linked administrative data to be used during survey processing as contextual reference information to identify survey responses that appear inconsistent, implausible or infeasible, thereby allowing likely measurement errors to be flagged. For example, a survey respondent may report that they were unemployed throughout a given reference period, but linked PAYE records indicate earnings at the beginning of that period, suggesting a potential inconsistency arising from misunderstanding of the reference period or employment definition. For government-funded surveys, such linkage may be undertaken under a “public task” basis; in other contexts, respondent consent or another appropriate legal basis may be required.

The RDMF could also support improved handling of missing data. Where survey responses are missing for individual cases, these values could be imputed using information within the survey dataset, such as replacing the missing value with the mean or median, using regression models to predict the missing value, assigning missing values based on similar respondents (hot deck imputation), and multiple imputation to account for the uncertainty of the imputation. However, relying solely on survey data for imputation may limit the accuracy of imputed values, particularly when key predictors are unobserved or when missingness is related to characteristics that are poorly measured in the survey.

Provided respondents have given their consent, access to linked administrative data via the RDMF could provide a richer set of auxiliary variables to inform imputation models, potentially improving their predictive performance. For example, a wider range of predictive variables (e.g., occupation, location, tax band) could be incorporated into models used to impute missing income values. Because these auxiliary variables are observed for a large proportion

of the population, they may help reduce the risk of model misspecification and large imputation errors.

While the RDMF offers the above advantages, the use of administrative data to minimise measurement error requires careful consideration of data quality, definitional alignment, and temporal consistency, and does not remove the need to account for uncertainty.

## 7 Impact on non-statistical quality

While statistical quality is a fundamental aspect of survey quality, ensuring that data accurately reflects the true values and characteristics of the target population, it is not the sole determinant of a survey's overall utility and credibility. Total Survey Quality (TSQ) encompasses a broader set of criteria that include the needs and expectations of the users. These non-statistical dimensions play a crucial role in enhancing the overall value and applicability of survey data.

In this chapter, we delve into how the RDMF could improve survey quality, according to the user-driven dimensions of Total Survey Quality:

- **Relevance:** Using the RDMF could improve relevance by providing data that are more inclusive and granular than what would be produced using current survey methods. (section 7.1).
- **Timeliness:** The RDMF could improve timeliness by providing more accurate and up-to-date sampling frames. (section 7.2).
- **Credibility:** The RDMF could enhance credibility by enabling more effective fieldwork designs that increase response rates and reduce non-response bias. (section 7.3).
- **Completeness:** The RDMF could enhance completeness by providing access to a richer set of auxiliary variables, ensuring that datasets contain all necessary information for comprehensive analysis and interpretation (section 7.4).
- **Coherence and comparability:** The RDMF could enhance coherence and comparability by integrating administrative records with survey data, ensuring consistency and standardisation. However, careful management would be needed to address short-term transition challenges (section 7.5).
- **Accessibility and usability:** Using the RDMF in survey design and implementation is likely to involve stringent legal and ethical considerations affecting access, and technical and practical challenges affecting usability (section 7.6).

### 7.1 Relevance

Over recent years, we have observed significant changes in the requirements from data users and policymakers. There is growing awareness that data and evidence need to be more inclusive, ensuring "that everyone in society counts and is counted and no one is left behind." (UK Statistics Authority, 2022). Data users are also looking for more granular statistics in terms of geography (e.g., sub-national estimates), socio-demographics (e.g., more precise breakdowns of age, gender, and disability), and time (e.g., quarterly rather than annual estimates) (Government Statistical Service, 2021). However, traditional survey methods that rely on the Postcode Address File (PAF) as a sampling frame face challenges in delivering inclusive and granular data in a timely and cost-effective way. The RDMF offers a potential solution through its ability to integrate and link administrative records across departments.

*Inclusivity* in survey sampling is crucial for capturing a comprehensive and representative picture of the population. Traditional methods using the PAF often fall short in reaching certain demographic groups, such as rare populations, transient populations, individuals in communal establishments, or those without stable housing. The RDMF offers a transformative approach by enabling the integration of various administrative records to produce survey samples that are more inclusive (see section 5.2).

*Granularity* refers to the level of detail and specificity in the data, which is essential for addressing nuanced research questions and informing targeted policy interventions. Traditional survey methods, constrained by cost and logistical limitations, often struggle to provide highly granular data, particularly at small geographic levels or for specific subpopulations.

The RDMF could enhance the analytical value of survey data in two distinct ways. First, it would support fine-grained contextual analysis by enabling individual survey responses to be linked to highly granular administrative data describing local conditions—such as neighbourhood-level socioeconomic characteristics or service provision. This allows researchers to examine how outcomes vary by local context without implying statistically reliable survey estimates at very small geographic scales. Second, where survey coverage is limited and where appropriate methodological assumptions can be justified, linked administrative data could be used as auxiliary information to support model-based small area estimation. Such approaches could enable the production of estimates for neighbourhood or smaller geographies, subject to transparent methods, uncertainty quantification, and appropriate disclosure controls. Together, these capabilities would support both richer contextual understanding and, where suitable, more spatially granular estimates to inform targeted policy responses.

By enhancing both inclusivity and granularity, the RDMF could contribute to survey data that is more relevant and actionable for data users and policymakers, ultimately leading to better-informed decisions and more effective policies.

## **7.2 Timeliness**

In survey research, timeliness typically refers to the speed with which survey data are collected, processed, and released, such that results remain relevant for describing current conditions and informing decisions.

Timeliness could be improved for face-to-face surveys if the RDMF is used to produce more accurate and up-to-date sampling frames, for example through the use of the Location Index,

which is more up-to-date and offers more comprehensive coverage than the PAF for sampling residential addresses. In addition, administrative indicators such as occupancy status, utility usage, or tenancy records could be used to identify remaining ineligible addresses in advance. Together, these enhancements could reduce unproductive interviewer visits and marginally shorten the effective duration of face-to-face fieldwork. For other survey modes, improved frame accuracy and the integration of administrative data would not necessarily reduce the overall fieldwork period and may, in some cases, extend it through more targeted contact strategies, including sequential mixed-mode approaches. However, these approaches could still enhance fieldwork efficiency and resource allocation by focusing effort on higher-yield cases and improving response outcomes.

In terms of timeliness, the RDMF could support more predictable and controllable post-collection processing by facilitating automated validation and consistency checks that cross-reference survey responses with relevant administrative information. While such checks may initially identify a greater number of apparent inconsistencies requiring review, they could enable earlier identification and more targeted prioritisation of editing tasks, reducing late-stage revisions and rework. In addition, the use of more accurate sampling frames and better-informed fieldwork strategies could contribute to improved coverage and response patterns across key population groups. Although post-survey adjustment and weighting would remain necessary, these improvements could help refine existing weighting approaches and reduce their sensitivity to large or complex adjustments, supporting more stable estimates and a smoother progression to final outputs.

### **7.3 Credibility**

For survey data to be credible, they need to be considered trustworthy by the users of the data, particularly those relying on survey findings to inform policy and decision-making. In social surveys, credibility is increasingly affected by declining response rates, the risk that some groups are systematically under-represented, and reliance on complex adjustment methods that can be difficult for users to interpret or assess.

The RDMF could help address these concerns by providing a transparent and governed framework for the use of administrative data in sampling, targeted fieldwork, validation, and post-survey adjustment. More systematic use of administrative data to inform sampling and fieldwork strategies could contribute to higher participation and help mitigate risks of non-response bias. Separately, clear documentation and consistency in how these methods are applied would strengthen user confidence that survey estimates are produced using robust, impartial, and well-understood processes.

Looking ahead, although not currently possible within existing legal and governance arrangements, a future approach that permitted the controlled linkage of names to sampled units could support more personalised contact strategies and improve response rates. While higher response rates do not in themselves guarantee reduced non-response bias, persistently low response rates are widely perceived by data users as a threat to survey quality. Improvements in response rates could therefore strengthen users' trust and confidence in the credibility of survey estimates, provided such approaches were implemented transparently and with appropriate ethical and privacy safeguards.

Administrative data accessed through the RDMF could potentially strengthen the credibility of survey outputs by enabling more robust and informative benchmarking against external population data. While some comparisons with administrative sources are already possible, the RDMF could allow surveys to draw on a wider range of administrative benchmarks that are more closely aligned with survey populations and geographies. This could provide policy users with greater reassurance that survey estimates have been assessed against other trusted evidence sources, and that any differences are understood and explained. Although such benchmarking cannot demonstrate the absence of bias, it offers a clearer and more credible basis for judging whether survey results provide a balanced and reliable picture of the population

#### **7.4 Completeness**

Completeness refers to the extent to which all required data are included in the dataset. This means that the dataset contains all the necessary information needed to meet the users' requirements, ensuring there are no significant gaps or missing data points that could affect analysis and interpretation. For instance, in a survey dataset, completeness would imply that all questions have been answered and all relevant variables are covered.

As discussed in section 6.4, the RDMF could provide access to a richer set of auxiliary variables for imputing missing answers to questions. For example, a wide range of predictive variables, such as occupation, location, and tax band, could be incorporated into models to predict missing income. Additionally, key information from administrative records could be added to the survey dataset, providing data that respondents might not be able to supply, such as detailed tax information.

#### **7.5 Coherence and comparability**

*Coherence* refers to the consistency and reliability of data derived from different methodologies on the same topic, ensuring they can be logically integrated to provide a comprehensive understanding. *Comparability* is the ability to reliably examine data across

various dimensions such as time, geography, and demographics, allowing for meaningful comparisons to identify trends, patterns, or differences. Overall, coherence ensures consistency across various data sources, while comparability ensures that data can be reliably compared across different contexts.

Integrating administrative records with survey data through the RDMF could support improvements in data coherence by enabling more systematic comparison across sources. Rather than assuming consistency, such integration would allow differences between survey and administrative data to be identified, assessed, and better understood. For example, demographic characteristics collected through surveys could be compared with corresponding administrative information to highlight areas of alignment or divergence, prompting further investigation of data quality, definitional differences, or timing effects. While the accuracy and suitability of administrative data may vary across sources, the RDMF would provide a structured framework for evaluating these differences, thereby supporting more informed interpretation and improved overall coherence of the evidence base.

Linking survey data with administrative records could also improve comparability across datasets. Administrative records often provide standardised variables that can serve as common reference points, making it easier to compare survey data across time periods, geographic areas, and demographic groups. In addition, the RDMF supports linking data once using a single, approved methodology and enabling those linked datasets to be reused for multiple purposes. This reduces the risk that different teams apply slightly different linkage methods or assumptions, which could otherwise lead to inconsistent results. By promoting consistent linkage processes alongside the use of common benchmarks, the RDMF helps ensure that comparisons across datasets and over time are more reliable and methodologically coherent.

While switching from the PAF to the RDMF could improve the coherence and comparability of survey data, it may introduce short-term challenges for continuous and longitudinal surveys regarding trend continuity and comparability. For example, improved coverage might result in a sample that better reflects the inferential population, but it may not be demographically comparable to data previously collected using a PAF sample. Similarly, replacing survey measures of income with administrative indicators could reduce measurement error, but it may affect temporal comparability, making it unclear whether differences in annual estimates before and after RDMF introduction are real or artefactual. To mitigate these negative impacts, careful management of the transition would be essential. This includes detailed documentation, calibration of historical data, and pilot testing to ensure that trend analyses remain robust and meaningful.

## **7.6 Accessibility and usability**

Accessibility refers to the ease with which users can obtain and access survey data, while usability pertains to how easy it is for users to understand, navigate, and effectively use the survey data.

The RDMF could enhance accessibility by offering comprehensive, inclusive, and timely data from a centralised source, eliminating the need to gather information from multiple sources. Additionally, it could improve usability through the provision of its high-quality, detailed information and robust data integration. However, these benefits are likely to come with technical challenges, as well as stringent legal and ethical considerations, requiring careful navigation of permissions and compliance with data protection regulations (see section 8).

## 8 Constraints and possible solutions in using the RDMF to improve survey quality

This chapter addresses our second research question: What are the constraints and possible solutions for using the RDMF in the design and implementation of social surveys? While this paper so far has explored the potential for the RDMF to improve survey quality, it is important to understand the limitations that shape what is currently feasible.

Across both statistical and non-statistical dimensions of quality, a set of cross-cutting constraints point to foundational limitations that would need to be addressed before the RDMF could support social survey processes in a reliable and systematic way. These constraints do not affect all Core Indexes equally; in particular, the Location Index is the most mature and currently the only index suitable for social survey sampling applications.

Before considering these constraints and potential solutions, it is important to clarify several points about the current design and remit of the RDMF:

- The RDMF was created to support data linkage, not to act as a sampling frame or a source of person-level data. It enables datasets to be linked securely through unique identifiers but does not provide access to the underlying identifying information. Only address and geography data are openly available (through the Location Index, which exists independently of the RDMF as well as forming part of it). However, while this data can be accessed directly, it does not include the RDMF identifiers needed to link to other RDMF indexes.
- The RDMF does not provide population definitions (e.g. households, usual residents) and should not be considered equivalent to a population register. It includes only those individuals who appear in administrative systems within a given period.
- Use of the RDMF in surveys would require access permissions that are not currently in place. Existing data-sharing agreements allow ONS to link administrative data for statistical production and research, but do not automatically cover broader survey applications.
- The RDMF is a dynamic resource, updated on different refresh cycles. Any survey use would need to manage this versioning carefully.
- Developing an RDMF-based sampling resource would require significant new investment and system development. While the RDMF could theoretically support person-level sampling, this would require a fundamental shift away from household-level survey designs which would bring substantial survey operational and

methodological challenges. At present, household sampling frames, are more stable, better understood, and specifically designed for household survey operations hence using the Location Index and keeping the survey design at household level would be operationally easier and still enable linkage to the demographic index to inform the design.

With this context in mind, the sections below summarise the main cross-cutting constraints and outline possible approaches to addressing them.

### **8.1 Inclusivity and representativeness**

As discussed in section 7.1, there is an increasing expectation among policymakers and data users that statistics should be inclusive, ensuring that everyone “counts and is counted”.

Representation and inclusivity are shaped by the ability to contact sampled individuals and their willingness to take part in the survey. The RDMF could in principle be used to identify groups that are more difficult to contact and/or less likely to participate, and target additional fieldwork efforts toward them to improve their representation in the final dataset. It could also support oversampling of these groups; however, oversampling alone does not correct for non-response bias if respondents differ systematically from non-respondents in unobserved ways. For example, oversampling migrant groups may still yield achieved samples dominated by those with stable, formal employment. Nonetheless, the RDMF could be more efficient in attaining a desired sample for example by oversampling small population sub-groups of analytic interest to ensure sufficient absolute numbers for robust sub-group analysis.

While the RDMF offers the potential to broaden coverage by integrating multiple administrative sources, it also introduces challenges for inclusivity and representativeness if used directly as a sampling frame. The Demographic Index systematically under-covers groups with weak or inconsistent administrative footprints, such as recent migrants, people in temporary accommodation, and those experiencing homelessness. At the same time, it may retain out-of-scope cases, such as people who have died and people who have emigrated. This reflects not only the frequency with which the Demographic Index is updated (currently around 6 to 12 months), but also the timeliness of updates to the underlying administrative sources. Together, these issues lead to simultaneous under- and over-coverage. This has implications not only for statistical quality (e.g. persistent bias even after weighting) but also for non-statistical dimensions such as trust, interpretability, relevance and comparability.

Person-level stratification would require the Demographic Index. Many aspects of the quality of the Demographic Index are well understood and positive, such as timeliness, accessibility and scalability, and work is ongoing to understand more about the accuracy of the index and

linkage using the index. However, quality requirements will differ depending on use, and further work is necessary to establish transparent standards and agreed requirements for survey-specific use cases. Without a clear quality benchmark, confidence in current suitability for reliably supporting individual-level sampling is limited. However, the Location Index provides full UK address coverage, including temporary accommodation and communal establishments, and can support address-level stratification.

A possible short-term approach is:

- Use the Location Index as the basis for sampling frames, given its maturity and accessibility.
- Introduce Demographic Index-based person information to inform household-level sampling only after applying strict quality checks and transparent eligibility rules.
- Publish coverage scorecards for the RDMF indexes to support transparency and user understanding.

In the longer-term, ongoing ONS work to better understand and improve the quality of the Demographic Index and future work to develop a population spine may, over time, enable more sophisticated sampling applications.

## **8.2 Quality and transparency of linkage**

A major constraint is the current lack of quantified evidence on linkage accuracy, particularly for the Demographic Index. The accuracy of the links, both within the RDMF and to external datasets, has not yet been fully assessed. This limitation reflects the novelty and complexity of the architecture rather than an assumption of poor quality. However, there are efforts currently in progress to unravel this complexity:

Ongoing work includes comparisons of linked datasets produced through traditional methods and through the RDMF, and the development of new accuracy metrics for composite data, including analyses of false-positive clusters (records that were wrongly matched together) and false-negative clusters (records that should have been linked but were not). However, linkage can follow multiple pathways, and each requires its own validation. This adds to the resource constraints which means this work will take time.

Currently, analysts do not receive end-to-end linkage diagnostics, limiting their ability to assess reliability, quantify potential bias, or understand subgroup-specific linkage error. Also, errors in early linkages persist through updates, and survey identifiers (e.g. incomplete names, inconsistent dates of birth) add further challenges.

These issues affect several components of survey operations:

- Adaptive and responsive designs may redirect effort in ways that are suboptimal if error in RDMF-derived indicators, such as mover status or predicted non-response, is large enough to limit their value for guiding decisions.
- Design weights, deduplication and sample control become less reliable when linkage error leads to duplication, misassigned addresses or missing matches.
- Specification, coverage and sampling error increase if auxiliary variables (e.g. age, benefits status and other characteristics) are noisy due to linkage uncertainty.

Non-statistical quality is also affected, as weak visibility of linkage accuracy reduces credibility, interpretability and coherence.

Possible solutions include:

- Complete current research assessing the quality of the Demographic Index and publish clear findings.
- Develop quantifiable linkage metrics (including false-positive and false-negative rates) and undertake sensitivity analyses to understand the implications of different match-rate scenarios.

### **8.3 Conceptual and definitional misalignment**

Administrative data are designed for service delivery, not for statistical measurement. As a result, administrative concepts often diverge from survey concepts. For example, Pay As Your Earn (PAYE) records from HM Revenue & Customs capture only formal payroll employment, whereas the International Labour Organisation (ILO) definition of employment includes self-employed people and those working minimal hours. Using PAYE alone as a proxy for employment would therefore lead to systematic under-counting of informal or irregular workers when compared to official employment statistics. Understanding the differences in definitions between administrative data and survey data and the implications of these differences is therefore important to consider when using administrative data in survey design and processes.

The RDMF does not define statistical concepts (e.g. households, usual residents) and does not provide mechanisms for reconciling definitional differences across sources. Survey teams would remain responsible for identifying and managing these inconsistencies.

Conceptual misalignment affects:

- Specification and coverage error, especially when administrative definitions of residents or households differ across sources or update on different schedules.
- Measurement validity, if administrative variables are substituted for survey concepts.

- Stratification and weighting, when misaligned constructs lead to or mis-specified control totals or noisy stratifiers.
- Temporal comparability, when administrative “as of” dates do not align with survey reference periods.

Non-statistical quality dimensions, including credibility, coherence and relevance, are similarly affected when users cannot determine whether differences reflect real change or definitional drift.

Possible solutions include:

- Developing definitional mapping between survey and administrative concepts.
- Introducing harmonisation frameworks for population alignment.
- Establishing source hierarchies, timestamping protocols and clear rules for resolving conflicts.
- Version control of the RDMF indexes used in each survey cycle.

#### **8.4 Quality of underlying administrative sources**

RDMF could support sampling strategies to improve the quality of survey estimates. However, the RDMF inherits the strengths and weaknesses of the administrative data from which it is built. Issues include outdated addresses, inconsistent identifiers, and missing information. Understanding these strengths and weaknesses, through quality metrics and high-quality meta data is important when using administrative data in any survey design or process.

Although the RDMF corrects some supplier errors, there is currently no mechanism for individual records to be corrected when mistakes such as these are found during analysis, as it is designed to represent the raw inputs. It can be difficult for analysts to disentangle different types of errors and understand their implications. Machine-learning methods may support anomaly detection, but these tools are not currently embedded within the RDMF.

Weak administrative inputs can lead to:

- Specification and coverage error, where the Demographic Index includes service users rather than the whole population.
- Stratification and sampling error, where auxiliary variables are noisy, outdated or inconsistently defined.
- Measurement, processing and estimation error, when administrative data contain inaccuracies.

Input quality issues also undermine credibility, comparability, relevance, usability and timeliness.

Possible solutions include:

- Strengthening feedback loops with data suppliers
- Developing quality dashboards summarising input quality and update cycles
- Continuing ONS research on the suitability of administrative data for supporting survey processes
- Producing high-quality meta data on the administrative sources

## **8.5 Governance, permissions, and legal constraints**

Governance and legal frameworks strongly shape the extent to which the RDMF can support survey processes. Although the RDMF complies with data protection legislation and uses privacy-preserving linkage methods, these safeguards mean that access is tightly controlled.

Data-sharing agreements with suppliers currently permit ONS to use the data for statistical purposes, but the specific permissions within existing agreements would need to be assessed and broader permissions negotiated for any intended uses that aren't captured. Even where suppliers support use in principle, internal security approvals may delay or prevent access. Delays may force survey teams to rely on outdated frames or weaker auxiliary variables, reducing timeliness and relevance.

In addition to re-negotiating data-sharing agreements with administrative data suppliers, many survey applications of the RDMF in non-government surveys would require explicit respondent consent to link survey responses to administrative records. Consent requirements therefore act as both a legal and methodological constraint, influencing sample composition, data availability, and analytic strategies.

Governance constraints affect:

- Statistical quality, by limiting access to person-level variables needed for stratification, non-response adjustment and coverage correction.
- Non-statistical quality, including credibility, usability, relevance and timeliness.
- Coherence and reconciliation, as restricted access makes it difficult to validate conflicts or examine coverage across sources.

Possible solutions include:

- Strengthening cross-government mandates for consistent data access.
- Expanding platforms that support controlled access to linked data.
- Engaging security and governance teams early to streamline the approvals process.

Public acceptability is also a critical component of the governance landscape. As discussed in Appendix B, public support for the use of administrative data in research is conditional on transparency, clear public value and strong safeguards. Even if legal permissions could be expanded, survey applications that rely on person-level administrative data would need to demonstrate clear public benefit and maintain trust.

## **8.6 Resourcing, skills and capacity constraints**

The ability of the RDMF to support survey processes also depends on the capacity to maintain and develop it. Current progress, particularly on the Demographic Index, is constrained by limited clerical, data engineering and analytical resource.

Clerical capacity is needed not only to review and resolve complex linkage cases such as conflicting addresses, duplicate entries or unclear matches, but also to support wider quality assessment and the improvement of linkage methods for large or complex clusters (linked administrative data believed to be associated with one person). Limited capacity allows inconsistencies to accumulate. Data engineering and analytical capacity is needed to develop linkage diagnostics, provenance tracking, conflict-resolution pipelines, and scalable validation tools. These capabilities help users to understand the strengths and limitations of the data, enabling more informed and proportionate use.

The RDMF currently provides a linkage backbone, not a population spine, sampling frame or harmonised household dataset. Additional development would be required to support survey operations end-to-end. Resource constraints also slow updates to documentation and metadata, affecting usability and transparency.

Possible solutions include:

- Investment in specialist data scientists, linkage engineers and survey methodologists.
- Expanding clerical support for quality assurance of the Demographic Index.
- Developing survey-specific pipelines and/or investment in further ONS research on products such as the Statistical Population Dataset (Office for National Statistics, 2023b) and Admin-Based Living Arrangements Dataset (Office for National Statistics, 2023a) that could support survey-specific uses of the RDMF.

## 9 Conclusion

This study set out to explore how, in theory, the RDMF could be used to improve the design and delivery of social surveys in the UK, and what limits currently stand in the way. Using a Total Survey Quality (TSQ) perspective allows us to assess not only statistical accuracy and precision, but also broader dimensions of quality such as relevance, credibility, coherence, and usability across the full survey lifecycle.

Overall, the analysis suggests that the RDMF has considerable potential to strengthen survey quality, but that this potential is uneven and conditional. Some applications are already realistic and relatively low risk, while others would require significant further development, clearer evidence on quality, and changes to governance and resourcing before they could be implemented responsibly.

### *Improving the end-to-end survey process*

In relation to the first research question, the strongest and most immediate opportunities lie in the use of the Location Index as a sampling resource. Compared with the small-user PAF, the Location Index offers more complete and timely coverage of addresses, better identification of ineligible cases, and clearer visibility of communal and institutional settings. These features support better alignment between target and inferential populations, reduce coverage error, and make design decisions around inclusion and exclusion more explicit.

Beyond frame construction, the RDMF could support better sample design and fieldwork strategies. Where reliable auxiliary information is available, it could be used to improve stratification, support targeted oversampling, and inform tailored fieldwork approaches aimed at mitigating non-response error. In the longer term, the possibility of selecting individuals rather than households—either directly from the Demographic Index or by linking sampled addresses from the Location Index to the Demographic Index—could remove the need for within-household selection methods that are prone to error and non-compliance, particularly in self-administered surveys.

The RDMF could also improve weighting in two important ways. First, by reducing coverage and non-response issues earlier in the survey process, it could reduce the extent to which weights need to compensate for representation errors, improving precision and effective sample size. Second, it could support richer and more timely weighting models by providing access to individual-level auxiliary variables that are better aligned with both survey participation and key outcomes than the area-level data typically used at present.

On the measurement side, the RDMF should be seen as a complement to surveys rather than a replacement. Its main value lies in strengthening how constructs are measured by combining

survey responses with administrative indicators where each source is best suited. This approach can improve measurement validity, reduce reliance on error-prone self-reports, shorten questionnaires, and lower respondent burden. Linked administrative data can also support more effective questionnaire testing, validation, editing, and imputation, while recognising that administrative sources are themselves imperfect and need to be used carefully.

### *Constraints and conditions for use*

Turning to the second research question, the review also makes clear that there are significant constraints on using the RDMF for social surveys at present. These constraints cut across data quality, linkage accuracy, conceptual alignment, governance, capacity, and public acceptability.

A key limitation is that the RDMF was built to support secure data linkage, not to act as a population register or a survey sampling frame. In particular, the Demographic Index does not yet have transparent standards or agreed requirements for survey-specific use cases to have sufficient confidence in using it to reliably support routine person-level sampling or stratification. Some groups are systematically under-represented in administrative data, while others may remain in the system longer than they should, leading to simultaneous under- and over-coverage. Although ongoing work within ONS will provide a clearer picture over time, these issues are currently difficult to quantify and manage, and they affect not only statistical quality but also credibility and interpretability.

Conceptual differences between administrative data and survey definitions present another major challenge. Administrative records are designed for operational purposes and do not map cleanly onto statistical concepts such as employment, households, or usual residence. Without explicit mapping, harmonisation and version control, substituting administrative variables for survey measures risks undermining measurement validity and comparability over time.

Governance and legal constraints further limit what is currently feasible. While data-sharing agreements currently permit ONS to use the data for statistical purposes, approval processes for any broader use not already covered could be slow and fragmented. These restrictions affect the timeliness and scope of possible applications and are closely tied to issues of public trust. Any expansion of RDMF use in surveys would need to clearly demonstrate public value and be accompanied by strong ethical and transparency safeguards.

In addition to re-negotiating data-sharing agreements with RDMF data providers, informed respondent consent would be required where responses from non-government surveys are linked at the individual level to administrative data, whereas government surveys typically rely

on 'public task' as the legal basis for such linkage. This represents a key practical and methodological constraint for non-government surveys.

Finally, realising the RDMF's potential for surveys would require sustained investment in skills, systems and capacity. At present, the RDMF provides a strong linkage backbone but not a survey-ready infrastructure. Developing coverage assessments, linkage diagnostics, quality dashboards and survey-specific pipelines would require coordinated effort across methodological, analytical and engineering teams.

### *Overall assessment*

In summary, this study finds that the RDMF could play an important role in improving the quality of social surveys in the UK, but that its benefits are neither automatic nor uniform. In the short term, the most practical and defensible uses lie in improved address-based sampling, better use of auxiliary information for weighting and fieldwork design, and selective support for measurement validation. More ambitious uses, particularly those involving person-level sampling or widespread replacement of survey measures, remain longer-term prospects that depend on further evidence, clearer governance arrangements, and broader public engagement.

Crucially, the RDMF should not be viewed as a substitute for surveys. Surveys remain essential for capturing attitudes, experiences and behaviours that administrative systems do not observe. The RDMF's value lies in helping surveys to better define who they represent, to reach them more efficiently, and to produce results that are easier to interpret, explain and trust.

## REFERENCES

- Biemer, P. P. (2010). Total survey error: design, implementation, and evaluation. *Public Opinion Quarterly* 74(5), 817-848.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design - For Market Research, Political Polls, and Social Health Questionnaire*. San Francisco, CA: Jossey-Bass.
- Cabrera-Álvarez P., Jäckle A., Moore J. C., Sakshaug J. W., Durrant G., Burton J. & Smith P. W. F. (2025) Are we stepping into the future? Exploring the representativeness of web-only surveys of the general population, *Survey Futures Working Paper* no. 7. Colchester, UK: University of Essex. Available at <https://surveyfutures.net/working-papers/>.
- Corlett, A. (2024). *Get Britain's Stats Working: Explaining alternatives to Labour Force Survey estimates*. Resolution Foundation Briefing.
- Couper, M. P. (2017). New Developments in Survey Data Collection. *Annual Review of Sociology* 43(1), 121-145.
- d'Ardenne, J., Bull, R., Das, A., Perera, Z. M., & Sexton, O. (2025). How to mitigate against measurement effects when surveys move online: A Measurement Effect Risk Framework (MERF) and web questionnaire guidance, *Survey Futures Survey Practice Guide* no. 2. Available at Colchester, UK: University of Essex. Available at <https://surveyfutures.net/practice-guides/>.
- Department for Transport. (2025). *National Travel Survey 2024: Technical Report*. Available at <https://www.gov.uk/government/statistics/national-travel-survey-2024-technical-report>.
- Department for Work & Pensions. (2025). *Deep material poverty: Financial year ending 2024*. Available at <https://www.gov.uk/government/statistics/deep-material-poverty-financial-year-ending-2024>.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th ed.). Wiley.
- European Statistical System (ESS). (2019). *Quality Assurance Framework of the European Statistical System*. Brussels: European Statistical System. Available at <https://ec.europa.eu/eurostat/web/quality/european-quality-standards/quality-assurance-framework>.

- Foster K. (1994) The coverage of the Postcode Address file as a sampling frame. Office of Population Censuses and Surveys, *Survey Methodology Bulletin* 34, 9-18. Available at <https://webarchive.nationalarchives.gov.uk/ukgwa/20160108193914/http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/smb-34/index.html>.
- Groves, R. M., & Lyberg, L. (2010). Total survey error: past, present, and future. *The Public Opinion Quarterly* 74(5), 849-879.
- Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5(3), 213-236.
- Lynn, P. (2024). A framework for identifying and addressing the risks of exclusion from social surveys, *Survey Futures Working Paper* no. 3. Colchester, UK: University of Essex. Available at <https://surveyfutures.net/working-papers/>.
- Maslovskaya, O., Lynn, P., Calderwood, L., Durrant, G., Fitzgerald, R., Nicolaas, G., & Williams, J. (2025). Survey Futures Position Statement on Response Rates: Supporting Material. *Survey Futures Position Statements*. Available at <https://surveyfutures.net/publications/position-statements/>.
- Organisation for Economic Co-operation and Development (OECD). (2011). *Quality framework and guidelines for OECD statistical activities, Version 2011/1*. Paris: Organisation for Economic Co-operation and Development. Available at <https://webarchive-storage.oecd.org/aemint-web-archive-prod/web-archive/de/de19cadf6a239b1927cdb0a0c866293288f90afbd05e355630e773e4d0504b3f.pdf>.
- Office for National Statistics. (2020). *Valuation Office Agency property attribute data: quality assurance of administrative data used in Census 2021*. July 2020. Available at <https://www.ons.gov.uk/peoplepopulationandcommunity/housing/methodologies/valuationofficeagencypropertyattributedataqualityassuranceofadministrativedatausedincensus2021>.
- Office for National Statistics. (2021). *Reference Data Management Framework: A tool and set of products used by the Office for National Statistics to safely link data together so that we can do more useful analysis*. Available at <https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/referencedatamanagementframework>.
- Office for National Statistics. (2023a). *Case studies for the population and migration statistics transformation: England and Wales, 2023*. June 2023. Available at

<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/casestudiesforthepopulationandmigrationstatisticstransformation/englandandwales2023>.

Office for National Statistics. (2023b). *Developing Statistical Population Datasets, England and Wales: 2021. July 2023*. Available at <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingstatisticalpopulationdatasetsenglandandwales/2021>.

Office for National Statistics. (2025). *Labour Force Survey quality update: September 2025*. Available at <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/labourforcesurveyqualityupdate/september2025>

Office for National Statistics. (2026). Survey Futures Reference Data Management Framework project - what is the RDMF: *March 2026*. Available at <https://www.ons.gov.uk/file?uri=/aboutus/whatwedo/programmesandprojects/referencedatamanagementframework/referencedatamanagementframeworkwhattistherdmfv03.pdf>.

Office for Statistics Regulation (2025). *Code of Practice for Statistics*. London: Office for Statistics Regulation. Available at <https://code.statisticsauthority.gov.uk/>.

Olson, K., Wagner, J., & Anderson, R. (2020). Survey Costs: Where are We and What is the Way Forward? *Journal of Survey Statistics and Methodology* 9(5), 921-942.

O'Toole, T., Cernat, A., Tzavidis, N., & Sakshaug, J. (2025). Data Integration: Options for Integrating Survey and Non-Survey Data. *Survey Futures Survey Practice Guide* no. 1. Available at <https://surveyfutures.net/practice-guides/>.

Reece, N., & Lynn, P. (2025, February 7). Assessing Methods for Within-Household Selection in Self-Administered Push-to-Web Surveys. Presented at ESS and NatCen Social Research survey methodology webinar series. Available at [https://www.youtube.com/watch?v=D8h\\_D4qug5k](https://www.youtube.com/watch?v=D8h_D4qug5k).

Smyth, J., Olson, K. M., & Stange, M. (2019). Within-Household Selection Methods: A Critical Review and Experimental Examination. In P. J. Lavrakas, M. W. Traugott, C. Kennedy, A. L. Holbrook, E. de Leeuw, & B. T. West, *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment* (pp. 23-46). Hoboken, NJ: John Wiley & Sons, Inc.

Tipping, S., & Nicolaas, G. (2006). In search of a population sampling frame for UK postal surveys. *Survey Methodology Bulletin*, 58, 56-74. Available at

<https://webarchive.nationalarchives.gov.uk/ukgwa/20160106185738/http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/index.html>.

UK Statistics Authority. (2022, December 29). *Inclusive Data Taskforce recommendations report: Leaving no one behind – How can we be more inclusive in our data?* Retrieved from UK Statistics Authority:  
<https://uksa.statisticsauthority.gov.uk/publication/inclusive-data-taskforce-recommendations-report-leaving-no-one-behind-how-can-we-be-more-inclusive-in-our-data/>.

Valet, P., Adriaans, J., & Liebig, S. (2019). Comparing survey data and administrative records on gross earnings: nonreporting, misreporting, interviewer presence and earnings inequality. *Quality & Quantity* 53, 471-491.

Willis, G. B., & Tessler, J. T. (1999). *Questionnaire Appraisal System QAS-99*. Available at [https://www.researchgate.net/profile/Gordon-Willis/publication/267938670\\_Question\\_Appraisal\\_System\\_QAS-99\\_By/links/54b7b26a0cf2e68eb2803f6a/Question-Appraisal-System-QAS-99-By.pdf](https://www.researchgate.net/profile/Gordon-Willis/publication/267938670_Question_Appraisal_System_QAS-99_By/links/54b7b26a0cf2e68eb2803f6a/Question-Appraisal-System-QAS-99-By.pdf).

## **APPENDIX A: INTERNATIONAL USE OF POPULATION REGISTERS IN SOCIAL SURVEYS**

This appendix summarises insights from interviews ONS conducted with five European National Statistical Institutes (NSIs) - Iceland, Italy, the Netherlands, Spain and Sweden – on how they use population registers across the end-to-end survey process. These findings provide international context for considering the potential role of the RDMF in UK survey design.

### **1. Overview**

Across all five NSIs, population registers play a central role across many aspects of the survey process. Key uses include:

- drawing sampling frames
- replacing or validating survey questions
- supporting imputation and weighting

The NSIs interviewed now operate hybrid systems that combine register data with targeted surveys to collect information not available in administrative sources, for example, attitudes, expectations and subjective measures. Although legal, cultural and operational contexts vary, all countries reported clear benefits from register-based methods, alongside challenges such as timeliness of administrative updates, population mobility, and integrating multiple administrative sources.

### **2. How population registers support survey processes**

Population registers play a central role in strengthening survey design and operations. This section discusses some uses of the population registers across different stages of the end-to-end survey process.

#### **2.1 Sampling**

All NSIs interviewed use population registers as the primary sampling frame for social surveys. Benefits include:

- near-complete population coverage
- removal of ineligible units before fieldwork
- ability to stratify at detailed demographic levels
- identification of specific sub-groups of interest
- household and individual-level sampling depending on survey needs

Examples include:

- Iceland: Labour Force Survey samples are drawn directly from the National Register, which includes personal ID numbers for all people aged 16-89 years living in Iceland.
- Italy: Register-based sampling underpins both annual census surveys and targeted surveys of rare populations.
- Netherlands: Registers have replaced most surveys, with only variables lacking a register analogue collected directly.
- Spain: Registers underpin household frames; where individuals are sampled, the household record is used to locate them.
- Sweden: Sampling frames are routinely enriched by linking to education, taxation and other registers.

## **2.2 Fieldwork operations**

Population registers support more efficient and targeted fieldwork. For example:

- Italy and Spain use registers to flag households where residents have moved or died to reduce time spent trying to contact ineligible households.
- Spain uses register information to request telephone numbers from communication authorities, improving contact rates.

## **2.3 Question replacement and pre-populating**

All NSIs reported extensive replacement of factual survey questions with register data where possible. This reduces burden and increases accuracy. Commonly replaced variables include:

- demographic characteristics
- income and employment indicators
- education and qualifications

Subjective variables remain survey-based. In Italy, surveys are increasingly pre-filled with administrative data; corrections by respondents do not yet feed back into registers, however Italy plan to implement this mechanism soon.

## **2.4 Validation, editing and imputation**

Registers are widely used to cross-validate respondent answers and fill missing data. Use of the register data enables missing values to be imputed more accurately because of the rich set of auxiliary variables available for imputation models. Examples include:

- cross-checking employment and income variables (Italy, Sweden)
- verifying demographic characteristics (Iceland, Netherlands)
- imputing missing values using register data (Sweden)

- augmenting survey responses with register updates collected after the survey to study outcomes such as fertility (Italy)

## **2.5 Bias assessment and weighting**

Population registers provide rich auxiliary variables to diagnose and adjust for non-response and coverage errors, improving representativeness and reliability of survey estimates.

- Iceland compares Labour Force Survey respondents with the National Register, uses auxiliary register variables for non-response analysis and applies logistic regression-based propensity scores to produce non-response weights.
- Italy uses register variables for calibration and model-based estimation in its annual census.
- The Netherlands supplements surveys with register variables to improve weighting and reduce sampling error.
- Sweden uses calibration variables from linked registers.

This leads to improved accuracy and smaller design effects, particularly when auxiliary information closely correlates with key survey outcomes.

## **3. Challenges**

Despite their benefits, NSIs face common challenges when relying heavily on registers.

### **3.1 Coverage**

Coverage issues are the most common challenge, particularly those related to migration. Iceland and Spain face delays identifying emigrants, leading to temporary over-coverage. Sweden sometimes accepts missing values and produces statistics based only on individuals with available data. Italy applies “signs of life” checks to identify missing or excess population and routinely sends data back to municipalities for validation.

In Iceland, the increasing use of foreign mobile numbers amongst migrants, enabled by free EU roaming introduced in 2017, has contributed to rising non-contact rates as these numbers are not captured in national systems. Iceland’s National Register also experiences under-coverage (e.g. undeclared work) and over-coverage (limited incentives to de-register when leaving the country). To mitigate these issues, the NSI maintains a separate statistical register using signs of life indicators to improve accuracy.

### **3.2 Timeliness**

Timeliness varies across registers, with some updated daily and others (such as education, income and taxation data) sometimes lagging by months or even years. Delays in reporting moves, migration and changes in identification status can temporarily reduce accuracy, while reliance on fixed reference dates means register information may not match individuals' circumstances at the point of being surveyed. Late availability of key administrative sources also shortens the time available for survey preparation and integration, affecting the overall timeliness and consistency of survey processes.

### **3.3 Operational and legal complexity**

NSIs face a range of operational and legal challenges when using population registers in survey processes. These relate to differences in statutory access, data-sharing arrangements, and the complexity of integrating multiple administrative sources:

- The Netherlands NSI has strong statutory authority to access all government registers for official statistics, but no legal right to access datasets held by the private sector.
- Italy reported that some administrative sources are received late in the year, leaving less time for census and survey processing.
- Spain receives monthly updates from all municipalities, which must be consolidated, de-duplicated and quality checked before being used statistically.

Even where legal access exists, NSIs often rely on collaboration with register owners. The Netherlands noted that early phases of register integration required building voluntary cooperation and establishing metadata standards before mandatory access was introduced. Maintaining coherence across many linked datasets requires strong metadata, validation and governance.

### **3.4 Conceptual differences**

Definitions of key variables can vary between surveys and administrative sources. For example, when the Netherlands adopted the register-based definition of employment (all individuals living in the Netherlands, even if they work in another country), this led to a shift in key employment statistics from those based on the survey definition (individuals working in the Netherlands regardless of where they lived). Definitions can also differ across countries, making international harmonisation difficult.

## **4. Public acceptability**

Across the five countries, public acceptability of register use is generally high but shaped by cultural and legal context. Common themes include:

- Strong trust in NSIs – particularly in Iceland, Sweden and the Netherlands.

- Expectation of data use for public benefit – especially in Italy and Spain where register updates are linked to access to public services.
- Reassurance from legal safeguards – transparency requirements, independent oversight, and data protection laws support confidence.
- Limited awareness of administrative data use – many NSIs note that the public is not always explicitly informed of register use in statistics.

Overall, maintaining transparency and demonstrating societal value are key to sustaining trust.

## 5. Relevance for the RDMF

Although the UK does not have a population register, international experience shows how high-quality, linkable administrative data can enhance survey design and quality. Key lessons relevant to consideration of whether the RDMF could be used to improve survey quality in the UK include:

- Linked administrative data can strengthen coverage and weighting, improving representativeness through richer auxiliary variables and bias-adjustment methods.
- Replacing factual survey questions with administrative data can reduce respondent burden and improve accuracy, provided the linkage is timely, stable and well-documented.
- Adaptive fieldwork is more effective when supported by integrated data, such as demographic or socioeconomic indicators used to prioritise cases.
- “Signs of life” methods can help to address over-coverage in sampling frames, particularly those affected by migration or outdated address information.
- Clear legal frameworks and trusted governance enable data sharing and public confidence.
- Hybrid systems remain essential; even countries with mature register-based infrastructures still rely on targeted surveys for subjective, behavioural and attitudinal information that administrative data cannot provide.

These international experiences highlight both the opportunities and the practical considerations relevant to the UK as it explores the future role of RDMF-linked data in improving survey design and quality.

## APPENDIX B: PUBLIC ACCEPTABILITY OF ADMINISTRATIVE DATA USE IN RESEARCH AND IMPLICATIONS FOR THE RDMF

This appendix summarises evidence on public attitudes towards the use of administrative data in research, drawing on UK and international literature. It identifies the conditions under which people tend to support administrative data use and how the RDMF aligns with these expectations.

### 1. Key findings

ONS commissioned a nationally representative survey in England and Wales in 2024 to explore public attitudes to data sharing (not published externally). The research identified four broad personas with varying levels of trust and concern: Optimists and Enthusiasts, who are broadly comfortable with administrative data use, and Apprehensives and the Disengaged, who are more cautious or uncertain. Across all groups, trust in ONS was high, but concerns remained about the origin of data, how accuracy is maintained, and how risks of misuse are managed. Participants were more accepting of data use when reassured that information would be anonymised, handled transparently, and used responsibly. The study also emphasised the value of clear, targeted communication, as different groups respond to different types of reassurance.

A broader review by ADR UK (Trust, Security and Public Interest: Striking the Balance, 2020) found that public acceptability of administrative data use tends to depend on three core conditions:

- Clear public benefit: there is greater support when data is used to improve services, inform policy or deliver societal value.
- Strong privacy and security protections: including de-identification, restricted access and secure data environments.
- Robust governance and transparency: with clear oversight mechanisms and visible accountability.

Trust in institutions is closely linked to these conditions, and transparency is critical in maintaining that trust.

International research echoes these themes. Cognitive testing by the US Census Bureau (Research on Public Opinion of Administrative Records: Cognitive Testing Report, 2023) found that people are more trusting of local government than federal agencies and feel more comfortable sharing data when community benefits are evident. Participants were often unclear what data was already held, highlighting the need for plain, accessible communication. New Zealand provides a useful comparison through its Integrated Data Infrastructure (IDI), which has a similar purpose to the RDMF. Research commissioned by

Statistics New Zealand (Public attitudes to data integration, 2016) reported that public acceptability depends on strong safeguards such as strict access controls, full anonymisation, and clear evidence of public benefit. Participants generally viewed Statistics New Zealand as a trusted data custodian and saw the IDI as providing more current and reliable information than existing alternatives.

Earlier UK studies by ONS (The Census and Future Provision of Population Statistics in England and Wales: Public attitudes to the use of personal data for official statistics, 2014), Ipsos (Dialogue on Data, 2014) and Scottish Government (Public Acceptability of Cross-Sectoral Data Linkage, 2012) echo many of the same themes, despite being more than a decade old. Participants often had limited understanding of how government handles data but were generally supportive of administrative data use when it is anonymised, secure and demonstrably in the public interest. Privacy, confidentiality and transparency remained the main concerns.

## **2. Emerging themes**

Across the literature reviewed, there is broad but conditional public acceptance of using administrative data in research. Support tends to depend on six consistent themes:

- Low public understanding: limited awareness of how data is collected, processed and shared contributes to uncertainty and scepticism.
- Clear public benefit: acceptability increases when data use leads to tangible improvements in services, policy, or value for money.
- Privacy, anonymity and security: reassurance about confidentiality is essential, particularly where data is used beyond its original purpose.
- Governance and oversight: people expect strong regulatory frameworks and visible safeguards.
- Trust and institutional confidence: transparency and independent oversight help maintain trust in data-using institutions.
- Communication and engagement: clear, accessible explanations of data use, safeguards, and benefits build understanding and support.

## **3. Implications for the RDMF**

The evidence suggests that when administrative data use is transparent, well-governed and clearly beneficial, it is generally acceptable to the public. This section considers how the RDMF aligns with these expectations.

### **3.1 Public awareness and understanding**

While blogs, conference presentations, research publications, and a dedicated ONS webpage have helped raise awareness of the RDMF among specialist audiences, these channels are unlikely to have significantly improved wider public understanding. A nationwide communications campaign would be necessary to reach and engage the general public at scale. More plain-English materials and practical case studies demonstrating real-world benefits would help make the RDMF more tangible to the public. As examples of how the RDMF is used in practice emerge, these should be shared proactively to support wider understanding.

### **3.2 Public benefit**

The RDMF is underpinned by the principle of delivering public good. By enabling faster, more consistent data linkage, it supports the production of higher-quality statistics and more timely insights for policy making and service improvement. Ethical review processes ensure that only projects with clear public value use RDMF-linked data. Individual ONS programmes could enhance this further by communicating more clearly how their specific uses of the RDMF deliver benefits for society.

### **3.3 Privacy and security**

Privacy protection is central to public acceptability, and the RDMF has been designed accordingly. It uses strict de-identification, secure storage and tightly restricted access, with only a small number of trained and vetted data engineers handling identifiable information. Researchers receive only de-identified datasets. The framework follows the “Five Safes” model to minimise risk and operates under formal data-sharing agreements. This framework sets out a series of principles that enable data services to provide safe research access to data, covering safe data, safe projects, safe people, safe settings, and safe outputs. In the RDMF, directly identifiable personal data is removed as soon as linkage is complete, and there is a clear principle for operational separation between linkage teams and analytical teams. These safeguards align with what the public expects when administrative data are used.

### **3.4 Governance, oversight and accountability**

The RDMF benefits from extensive internal governance. Oversight is provided by the RDMF Oversight Board, methodological assurance panels, and the National Statistician’s Data Ethics Advisory Committee. A dedicated Data Linkage Ethics team provides additional scrutiny. Multiple Data Protection Impact Assessments (DPIAs) cover the RDMF’s design and processes, with annual reviews. New DPIAs will be produced as additional use cases emerge, ensuring continued accountability and compliance.

### **3.5 Trust and institutional confidence**

Public trust in ONS remains high, offering a strong foundation for acceptance of the RDMF. Evidence shows that people are more comfortable with data linkage when it is carried out by trusted institutions. Continued transparency, robust safeguards and clear demonstration of public benefit will help sustain this trust. Proactive communication would also provide opportunities to explain how the RDMF can strengthen the quality and reliability of statistics.

### **3.6 Communication and public engagement**

There is already substantial evidence, across the UK and internationally, on public views about use of administrative data. Many earlier concerns have been addressed through improved safeguards and clearer communication. As a result, commissioning new primary research specifically on public acceptability of administrative data use is unlikely to be necessary or proportionate. Instead, the priority should be to maintain transparent, accessible communication about how the RDMF works, what safeguards are in place, and how RDMF-enabled work delivers public benefit. Clear, consistent messaging and ongoing engagement will support continued public confidence as the RDMF becomes embedded across ONS programmes.

## **References**

ADR UK (2020). Trust, Security and Public Interest: Striking the Balance

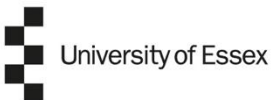
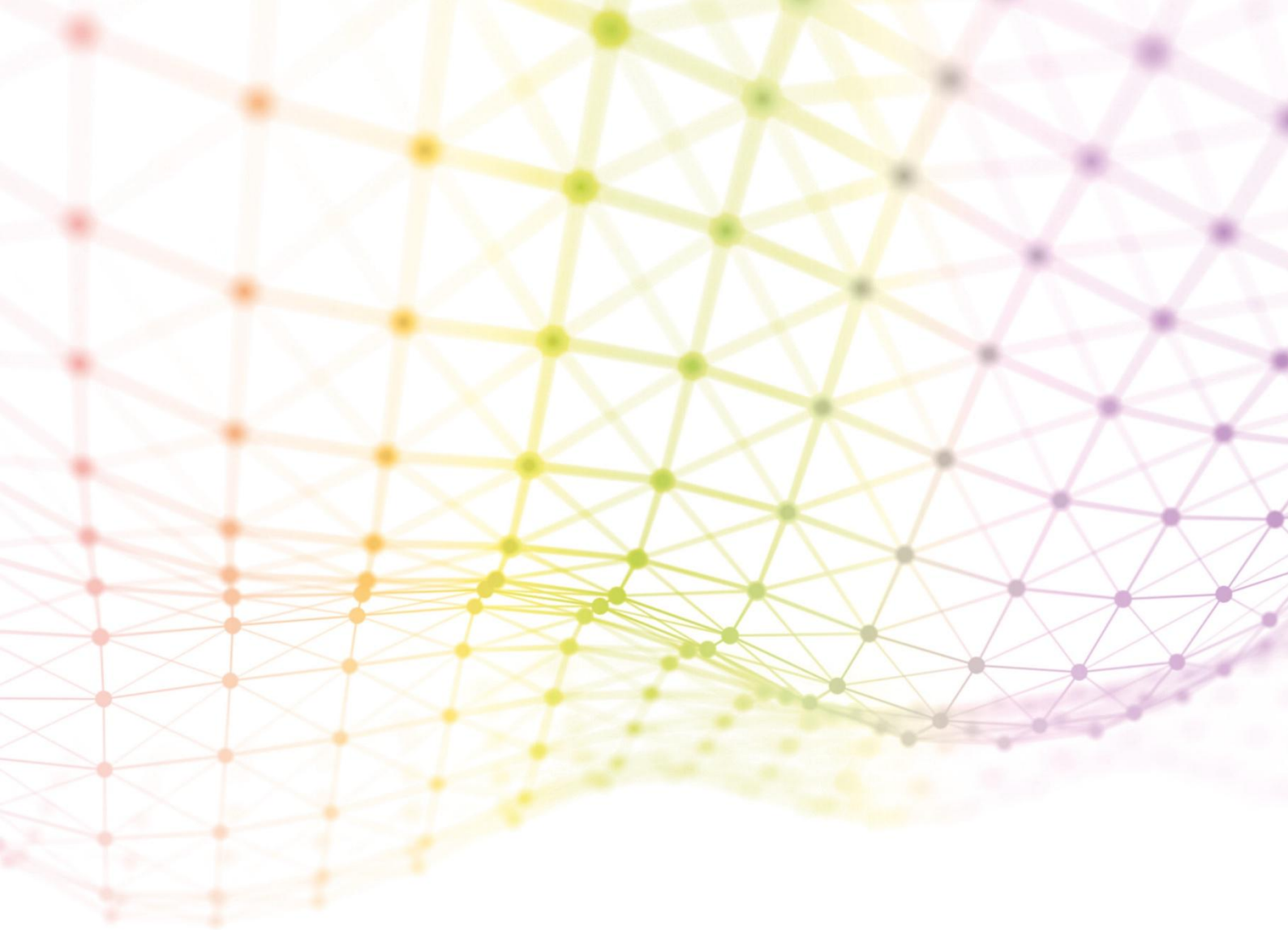
US Census Bureau (2023). Research on Public Opinion of Administrative Records: Cognitive Testing Report

Statistics New Zealand (2016). Public attitudes to data integration

ONS (2014). The Census and Future Provision of Population Statistics in England and Wales: Public attitudes to the use of personal data for official statistics

Cameron, D., Pope, S., Clement, M. (2014). Dialogue on Data: Exploring the public's views on using administrative data for research purposes

Scottish Government (2012). Public Acceptability of Cross-Sectoral Data Linkage: Deliberative Research Findings



[www.surveyfutures.net](http://www.surveyfutures.net)