WORKING PAPER

Quality of life

European Quality of Life Survey 2016: Quality Assessment

European Quality of Life Survey 2016

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

This report details the results of an independent quality assessment of the processes and outputs of the 4th round of the European Quality of Life Survey (2016), It first assesses survey process quality using a Quality Assurance Plan developed by Eurofound that includes a number of specific quality indicators linked to the quality framework and quality dimensions used by the European Statistical System (Relevance and Timeliness, Accuracy, Punctuality, Accessibility, and Coherence and Comparability), Next, it examines the quality of the main survey outputs including the final source questionnaire in English and the final dataset. The report then assesses adherence of the 4th EQLS to a set of best practices for 3MC (multinational, multiregional, and multicultural) surveys on the basis of recent methodological literature and examples of other comparable 3MC surveys in the European context. Based on a comprehensive review of the processes and outputs of the 4th EQLS, overall, Eurofound has followed design, implementation, and quality control and quality assurance best practices for 3MC surveys, and in some cases exceeded the quality metrics observed in comparable surveys. Recommendations are provided for improving the survey in the future.

Eurofound (2018), European Quality of Life Survey 2016: Quality Assessment. On-line working paper.

Authors: Julie de Jong, Kristen Cibelli Hibben (University of Michigan)

Contributors: Beth-Ellen Pennell, Dan Zahs (University of Michigan); Brita Dorer (GESIS)

Research Manager/s: Tadas Leoncikas, Sophia MacGoris

Eurofound reference number: WPEF18059

Related report/s: Eurofound (2018), European Quality of Life Survey 2016: Technical and fieldwork report. On-line working paper (WPEF18016)

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency, whose role is to provide knowledge in the area of social, employment and work-related policies. Eurofound was established in 1975 by Council Regulation (EEC) No. 1365/75 to contribute to the planning and design of better living and working conditions in Europe.

European Foundation for the Improvement of Living and Working Conditions

Telephone: (+353 1) 204 31 00

Email: information@eurofound.europa.eu

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

Contents

Executive Summary

Purpose

The 4th European Quality of Life Survey (EQLS 2016), fielded from September 2016 to March 2017 surveyed nearly 37,000 people across 33 countries – the 28 EU Member States and 5 EU candidate countries. A team of survey methodologists, specialising in cross-cultural survey research at the University of Michigan, was awarded a competitive contract to conduct an independent quality assessment of the processes and outputs of the EQLS 2016 and to make recommendations for improving future surveys. The report details the results of this quality assessment of the EQLS 2016 and subsequent recommendations.

Methodology of the quality assessment

There are several approaches to assessing survey quality, including: 1) total survey error; 2) fitness for intended use; and 3) monitoring survey production process quality. The assessment of the 4th EQLS presented in this report draws upon each of these approaches in an integrated framework.

We first assess survey process quality using a Quality Assurance Plan (QAP) developed by Eurofound prior to the release of the tender for the 4th EQLS. The QAP consists of a detailed set of quality indicators and associated targets, grouped by main survey lifecycle stage and linked to the quality dimensions widely used within the European Statistical System as a framework to assess quality (European Union, 2015), Our assessment focuses primarily on indicators related to the dimension of accuracy, which is essential for the other quality dimensions to be relevant (Biemer & Lyberg, 2003), and briefly reviews compliance with indicators targeting the other quality dimensions.

We next examine the quality of the key outputs from the survey, including the final source questionnaire in English and the final dataset. We assess the former using a set of criteria drawn from widely-used guidelines and principles for questionnaire design. Key outputs evaluated in the dataset include but are not limited to final sample disposition rates (e.g., response, noncontact, etc.) and final sample composition at the country level (e.g., distribution of sample by gender, age, etc.), Both disposition and composition rates from the 4th EQLS are compared to the 3rd EQLS (completed in 2012), as well as to recent rounds of the European Social Survey (ESS),

We then assess adherence of the 4th EQLS to a set of best practices for other multinational, multiregional, or multicultural surveys (referred to as ‘3MC’ surveys), which we define based on the work of comparable 3MC surveys in the European context and the recent relevant methodological literature.

The assessment of survey processes

The assessment of survey processes is organised into the four main stages of the survey lifecycle: sampling frame development, questionnaire development, fieldwork, and weighting. Based on our evaluation of compliance with the QAP, focusing on the required accuracy-related indicators, we find a high level of compliance across all stages. While there are several exceptions, including noncompliance with a specific indicator related to the final source questionnaire, we assessed non-compliance as generally having a minimal effect on data quality. While we note ways that it could be improved, we also find that the development and use of the QAP is an important tool and an advance for assuring and controlling quality in 3MC surveys.

The assessment of survey outputs

Our assessment of the final source questionnaire identified some design decisions that could have implications for comparability across countries and waves of the EQLS and some specific items that a survey methodologist trained in 3MC questionnaire design may have identified for revision in an expert review. Comparisons of the sample composition of the 4th EQLS with data from Eurostat and the ESS suggest that the sampling and fieldwork processes are in line with other major cross-national data sources. Analysis of both the coefficient of variation and the design effects provide evidence of increased efficiency, which lessens the impact of subsequent weighting calculations on estimates of statistical precision, while increasing comparability. Response rates and contact rates improved in a number of countries compared to the 3rd EQLS. However, cooperation rates largely suffered in the 4th EQLS. The adoption of the AAPOR standard outcome codes for the 4th EQLS also facilitated the harmonisation of outcome codes and comparability to past EQLS waves and other 3MC surveys.

The assessment of adherence to 3MC survey best practices

Overall, a high level of standardisation and quality was achieved in sampling frame development and sampling procedures for the 4th EQLS, although there are areas for continuous process improvement, as discussed in the Recommendations Section. The questionnaire development process incorporated many best practices including consultation with subject matter experts, a translatability assessment, team translation, and a sizable investment in pretesting. However, the process would have benefited from additional expert review and pretesting, as well as more thorough documentation, as detailed in the Recommendations Section. Fieldwork is a particular area of strength for the 4th EQLS. Key fieldwork procedures were highly standardised including the respondent selection process at the household level, the use of a standardised CAPI instrument for both sample management and questionnaire administration, the mode of initial contact, the pilot test protocol, and number of call-backs (with very few exceptions), The Recommendations Section details strategies and suggested changes to fieldwork procedures to increase data quality in future surveys. While we note a few areas where documentation could be improved, we find overall, that changes in weighting procedures between the 3rd and 4th wave were positive and in line with best practices.

Overall findings

Based on a comprehensive review of the processes and outputs of the 4th EQLS, we find that Eurofound has generally followed design, implementation, and quality control and quality assurance best practices for 3MC surveys, and in some cases exceeded the quality metrics observed in comparable surveys.

Recommendations

We draw on the principles of the SWOT framework, which considers (s)trengths, (w)eaknesses, (o)pportunities, and (t)hreats, to prioritise recommendations, as follows, for the main stages of the survey lifecycle by cost and relative impact or trade-off in addressing key sources of survey error, using the Total Survey Error (TSE) framework.

Sampling frame development

- Consider alternate respondent selection methods
- Calculate effective sample size
- Clearly define the target population
- Consider alternatives to enumeration methods
- Thoroughly document sampling frame sources

Questionnaire development

- Develop research questions/aims and an analysis plan for new questions
- Involve a survey methodologist trained in 3MC questionnaire design
- Carry out an expert review
- Adopt a uniform approach to assessing final translations
- Produce show cards and standard protocol for their use
- Expand documentation of the questionnaire development process
- Consider using advance translation or a team approach to the translatability assessment
- Expand the use and documentation of cognitive interviewing

Fieldwork

- Thoroughly document the CAPI development & testing processes
- Conduct nonresponse bias analyses
- Standardise the mode of initial contact
- Implement partially interpenetrated fieldwork assignments (at least two interviewers should be assigned to each primary sampling unit to enable analysis of interviewer effects separately from effects of the geographical area)
- Implement responsive design techniques
- Consider an alternative interviewer pay structure

Weighting

- Investigate the effect of weighting changes
- Thoroughly document calibration data and information about the sources
- Provide additional documentation related to weighting

Also included are recommendations for Eurofound's reporting of output statistics, data dissemination, and data disclosure, and how the QAP could be enhanced in the future.

1. Introduction

1.1 Background of the 4th EQLS

As a substantive research topic, quality of life is a broad concept, with countless latent concepts. To understand the numerous components comprising the quality of life, the European Quality of Life Surveys (EQLS) are carried out every four to five years across all European Union (EU) Member States, and extended to EU candidate countries. These surveys collect unique and critical data from the public on current life conditions as well as their attitudes on a wide range of related topics, including employment, income, education, housing, family, health, work-life balance, happiness, life satisfaction, and perceptions of societal quality. Equally important, the survey instrument replicates a significant number of questionnaire items from wave to wave, permitting trend analysis in the cross-sectional datasets.

The data captured by the EQLS are used both in cross-national comparative analyses as well as for in-depth analyses in individual countries and can have important policy implications at both the national and EU level. For example, using data from the 3rd EQLS, an investigation of the relationship between mental well-being and levels of financial strain, and the interaction with different neighbourhood characteristics provided evidence that good access to green/recreational areas decreased the effects of financial strain on mental health (Mitchell et al., 2015), Another study using data from the 3rd EQLS examined the extent to which access to green/recreational areas moderates the negative impact of neighbourhood noise and air quality on health, finding that better access to green areas reduced rates of self-reported poor health associated with neighbourhood noise and poor air quality (Dimitrova & Dzhambov, 2017), These are but a couple of examples of analyses based on EQLS data, beyond Eurofound's own reports, with results that are directly relevant to public policy. For EQLS data to fuel such research, it is essential that the highest level of data quality is assured.

This report is a quality assessment of the 4th EQLS which surveyed nearly 37,000 people across 33 countries – the 28 EU Member States and 5 EU candidate countries. The questionnaire included items categorised primarily into three areas: quality of life, quality of society, and quality of public services. In March 2017, Eurofound published a tender for a comprehensive quality assessment of the 4th EQLS, to which a team of survey methodologists specializing in cross-cultural research at the University of Michigan applied and to whom the contract was subsequently awarded. This report is the result of our assessment of the overall quality of the 4th EQLS.

1.2 Approaches to survey quality in 3MC research

In the 4th EQLS, as in other multinational, multicultural, or multiregional surveys, referred to as '3MC' surveys, success hinges on the comparability or equivalence of data across many cultures and countries. Yet the challenges of documentation, survey quality assessment procedures and criteria are far more complex in this context.

The concept of *survey quality* is central to the entire survey lifecycle and therefore impacts all stages of the survey process. Given the magnified quality assurance and quality control challenges in 3MC surveys, it is of critical importance to include such quality programs in these surveys, insofar as possible, and as advanced by the 4th EQLS.

There are several approaches to assessing survey quality, including: 1) total survey error; 2) fitness for intended use; and 3) monitoring survey production process quality, which may be affected by survey infrastructure, costs, respondent and interviewer burden, and study design specifications. Hansen et al. (2016) provide an integrated approach that brings together these three approaches in a comprehensive framework. We review each of these components and then discuss how our assessment of the 4th EQLS draws upon this integrated framework.

Total survey error (TSE)

Total survey error (TSE) (Groves et al., 2009) is widely accepted as the organizing framework in the design and evaluation of single-country surveys and is increasingly being applied to 3MC surveys (Pennell et al., 2017), Errors in survey estimates consist of variances of estimates (reflecting estimate instability over conceptual replications) and systematic deviations from a target value ('biases'), TSE defines quality as the estimation and reduction of the mean square error (MSE) of statistics of interest, which is the sum of random errors (variance) and squared systematic errors (bias), even though the MSE for each individual statistic in a survey is not typically estimated (Vehovar et al., 2012), TSE takes into consideration both measurement (construct validity, measurement error, and processing error), i.e., how well survey questions measure the constructs of interest – as well as representation (coverage error, sampling error, nonresponse error, and adjustment error), i.e., whether one can generalise to the population of interest using sample survey data (Groves et al., 2009), In the TSE perspective, there are cost-error trade-offs, that is, there is tension between reducing these errors and the cost of doing so. Pennell et al. (2017), Smith (2011), and the Cross-Cultural Survey Guidelines (Survey Research Center, 2016) have expanded the traditional TSE framework to include the concept of 'comparison error.' Originally defined by Smith (2011), comparison error is the error introduced across each stage of a 3MC survey as well as the aggregate of error across all stages. The TSE framework, with the inclusion of the concept of comparison error, helps organise and identify error sources and estimates their relative magnitude, which can assist those planning 3MC surveys to evaluate design and implementation trade-offs.

Fitness for intended use

The TSE framework, which has been argued to lack a user perspective, can be supplemented by fitness for intended use (Biemer & Lyberg, 2003), Fitness for intended use is multidimensional and focuses on criteria for assessing quality in terms of the degree to which survey data meet user requirements. By focusing on fitness for intended use, study design strives to meet user requirements in terms of survey data accuracy and other dimensions of quality including comparability, relevance, accuracy, timeliness and punctuality, accessibility, interpretability, and coherence. In this perspective, ensuring quality on one dimension (comparability) may conflict with ensuring quality on another dimension (timeliness); and there may be tension meeting user needs in terms of both survey error and fitness for use. However, the overall aim is to optimise quality, minimise costs and burden, and recognise and document design constraints at all levels.

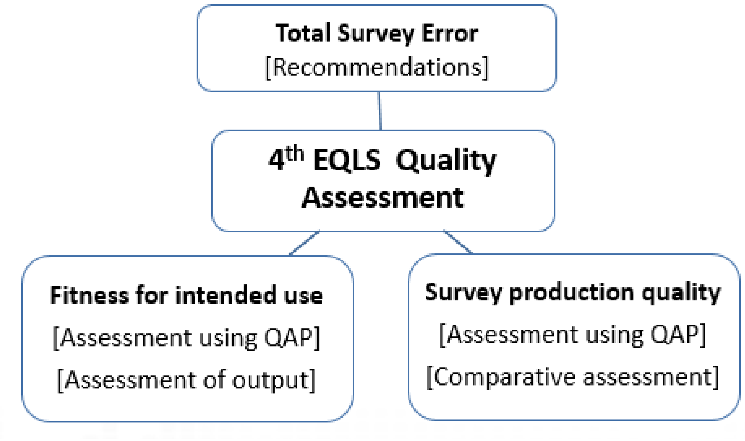
Monitoring survey production quality

Monitoring survey production process quality emphasises the notion of continuous process improvement (Groves et al., 2009), This approach focuses on quality at three levels: the product, the process, and the organisation (Lyberg & Biemer, 2008), The product quality, as mentioned by Lyberg and Stukel (2010) is the expected quality of survey deliverables, which is often decided by clients and/or data users. Process quality refers to the quality of the processes that generates the product. One way to monitor and control process quality is to choose, measure, and analyse process variables relevant to the particular survey (Lyberg & Stukel, 2010), A focus on survey production process quality requires the use of quality standards and the collection of standardised study metadata, question metadata, and process paradata (Couper, 1998), and is operationalised through the quality control process guided by quality planning and assurance. The quality control outcome measures are intended to result in a quality profile which can also be used to make recommendations for quality improvements, and subsequently reflected in future quality planning and assurance.

1.3 Approach to the 4th EQLS assessment

Each of the afore mentioned approaches to assessing survey quality has strengths and weaknesses. TSE by itself is not sufficient when thinking about survey quality because it lacks a user perspective. Biemer and Lyberg (2003) argue that the TSE framework should be supplemented with a multidimensional paradigm comprised of criteria for assessing quality in terms of the degree to which survey data meet user requirements; i.e., 'fitness for intended use'. Fitness for intended use lends an aspect of practicality and provides a general framework for assessing quality through several essential quality dimensions, and integrates TSE through the accuracy dimension. Lastly, survey process quality acknowledges the critical effect of processes on the end result. We draw an integrated approach for the assessment of the 4th EQLS, as illustrated in Figure 1, which shows how the components of the assessment are informed by the theoretical approaches to quality, as outlined above.

Figure 1. Integration of Quality Approaches



We begin our assessment in Section 2 (*Evaluation of survey processes per the 4th EQLS Quality Assurance Plan*) by evaluating survey process quality using the Quality Assurance Plan (QAP), which

Eurofound developed to accompany the Terms of Reference for this latest round. The QAP consists of a detailed set of quality indicators and associated targets grouped by twelve themes associated with the stages of the survey lifecycle. Each quality indicator is linked to one of the following quality dimensions: Relevance and Timeliness, Accuracy, Punctuality, Accessibility, and Coherence and Comparability. These criteria are widely used within the European Statistical System as a framework to assess quality (Eurostat, 2015), The QAP integrates two approaches to survey quality—fitness for intended use and monitoring survey production process quality—by monitoring and assessing quality according to each of the fitness for intended use quality dimensions. While the QAP separated the survey lifecycle into twelve themes, we have reorganised the indicators into the following four sections:

- Sampling frame development;
- Questionnaire development & advance translation, cognitive testing, and translation;
- Fieldwork (implementation, monitoring, contact procedures, nonresponse, and paradata); and
- Weighting.

Although the QAP indicators span five quality dimensions, our assessment in Section 2 focuses primarily on the indicators related to the dimension of accuracy, which is 'considered fundamental to product quality' (Biemer et al., 2014, p. 381), As cited further in Biemer et al. (2014), 'Biemer and Lyberg (2003) viewed accuracy as the dimension to be optimised in a survey while the other dimensions (the so-called user dimensions) can be treated as constraints during the design and implementation phases of production. They argued that sufficient accuracy is essential for the other quality dimensions to be relevant' (p. 386), We also briefly review compliance with indicators targeting the other quality dimensions. The section concludes with an evaluation of the current QAP as a framework for assessing quality in the 4th EQLS.

In Section 3 (*Assessment of output*), we examine the quality of the key outputs from the survey: the final source questionnaire in English and the final dataset. We assess the final source questionnaire using a set of predefined criteria drawn from relevant literature. Key outputs evaluated in the dataset include final sample disposition rates (response, refusal, cooperation, and contact rates) and final sample composition at the country level (e.g., distribution of sample by gender, age, etc.), Both disposition and composition rates from the 4th EQLS are compared to both the 3rd EQLS (completed in 2012), as well as to recent rounds of the European Social Survey (ESS), We also consider other types of output and analyses completed by Eurofound and its data collection contractor, Kantar Public, such as those examining item nonresponse, case duplication, and interviewer workload, as well as summary statistics produced concerning design effects and the coefficient of variation, which alongside evaluation of processes factor into a comprehensive assessment of the quality of the final dataset.

Until quite recently, the quality frameworks most-oft used in 3MC surveys were originally structured for single-country surveys; the industry lacked a suitable framework for quality assurance, control, and assessment to address the complexities in 3MC surveys. To supplement the current quality framework used in the 4th EQLS, in Section 4 (*Comparative assessment based on current best practices*) we define for each of the four phases of the survey lifecycle a set of 3MC survey best practice guidelines based on the work of other major 3MC surveys in the European context, as well as recent contributions to the methodological literature. Using these guidelines, we assess the extent to which the 4th EQLS achieves these standards.

Section 5 (*Summary findings*) summarises the main findings from the assessment drawing on our evaluation of the survey processes and outputs of the 4th EQLS and how these relate to best practices in the preceding sections, highlighting the many strengths of the 4th EQLS as well as noting the potential areas for quality improvement.

We conclude our assessment in Section 6 (*Recommendations*) with a view toward continuous quality improvement, wherein we use the principles of a SWOT analysis to prioritise opportunities for improving survey quality. Recommendations are organised and prioritised considering both cost and relative impact or trade-offs in addressing key sources of survey error, using the TSE framework. In this section, as requested by Eurofound, we also provide recommendations on other topics which are not easily categorised into a survey lifecycle stage. For example, the 4th EQLS included a supplementary web add-on following the main data collection, and we consider how advances in web surveys in other 3MC countries can inform such future innovations in the EQLS. We also include recommendations for Eurofound's reporting of output statistics, data dissemination, and data disclosure. Lastly, we provide recommendations for how the approach to quality, more generally, and how the QAP, more specifically, could be enhanced in the future.

Our assessment of the 4th EQLS differs in some ways from external assessments of prior waves of the EQLS, and from those of Eurofound's other 3MC survey targeted at individuals, the European Working Conditions Survey (EWCS), Changes between waves to Eurofound's surveys impact approaches to the assessments (e.g., this assessment is the first opportunity to assess the QAP developed prior to the 4th EQLS), Our approach in this assessment is to integrate key frameworks of survey quality assessment, considering some of the same elements evaluated in previous reports, along with our understanding of current best practice in 3MC research.

The EQLS is uniquely positioned in the 3MC survey landscape in terms of both the breadth of the populations covered and its questionnaire content, but faces limitations with regards to available funding, which limits both sample size and the application of resource-intensive approaches to quality control. The prioritisation scheme used in our recommendations can help guide Eurofound's design and implementation decisions as it strives to assess and improve the quality processes and outcomes of future EQLS and other Eurofound surveys.

Eurofound provided access to the quality assessment team to comprehensive documentation and materials of the survey preparation and implementation. Data users are invited to consult the specific technical reports that are either available on Eurofound website or can be made available on request:

- *European Quality of Life Survey 2016: Technical and fieldwork report (WPEF18016)*
- Sample evaluation, enumeration and weighting report
- Quality assurance report
- Translation report
- Coding report
- Data cleaning and editing report

2. Evaluation of survey processes per the 4th EQLS Quality Assurance Plan

The first step in our assessment is to evaluate survey process quality in the 4th EQLS by examining compliance with the Quality Assurance Plan (QAP) developed by Eurofound.

The research and consultancy firm, Kantar Public, was awarded the data collection contract for the 4th EQLS and served as the coordinating centre. Following fieldwork, Kantar Public produced a Quality Assurance Report (QAR), Using as evidence this report, along with other documentation from Eurofound, we assign compliance scores, defined as follows, to each quality indicator:

- **Target met** – the relevant target was met;
- **Target not met** – the relevant target was not met;
- **Outcome unknown** – compliance with the relevant target is unknown due to a lack of documentation or because the criteria was not specified in the indicator or the outcome is not measurable; and
- **Target no longer applicable (N/A)** – circumstances changed so that the quality indicator and associated target are no longer applicable.

In the QAP, indicators and associated targets were further classified in terms of what we refer to as 'achievable', with the following categories defined as the extent to which a target is considered:

- RQ: Required; that is, those targets that have to be achieved,
- RW: Real world; that is, those targets that can be achieved, and
- IW: Ideal world; that is, those targets that are unlikely to be achieved.

Throughout our assessment of compliance, we assume that if Kantar Public or Eurofound cite a specific document as evidence of a quality indicator's achievement (or otherwise) of a target, then we consider the said document to exist in the repository of project documentation. If the document is considered to be a valid source of information by Kantar Public and/or Eurofound, then we too consider it to be valid. Likewise, if a specific date is specified as evidence (e.g., the date of an interviewer training session), this is considered as valid evidence.

The use of the QAP to assess quality revealed some limitations in the framework itself. These limitations will be discussed in Section 2.5, as will recommendations for revisions for future use in Section 6.7. Prior to that discussion, we use the framework as designed, and assess actual impact on survey quality, regardless of compliance.

2.1 Sampling frame development

All participating countries in the 4th EQLS were required to implement a probability-based sample design, using a high quality sampling frame when available, as well as to use strategies to minimise nonresponse bias and limit clustering effects to the extent possible in order to minimise design effects and standard errors while maximizing efficiency. The QAP includes a number of indicators to assess elements relating to these requirements.

Table 1. Accuracy Indicators in Sampling Frame Development (Required)

Indicator	**Target**	**UM Assessment**
RQ: Training materials cover selection of respondent within household	Yes	Target met
RQ: 100% of enumerators that take part in enumeration training	100%	Target met
RQ: Country specific enumeration plan ensures random selection of respondents	100%	Target met
RQ: Enumeration is checked in at least 10% of the PSUs	100%	Target met
RQ: Follow up action taken if deviations from country specific enumeration plan observed	100%	Target met
RQ: Specified information on stratification variables included in the reference statistics	100%	Target met
RQ: Sample size >= 1000	100%	Target met
RQ: Net sample size >= planned sample size	100%	Target met

The quality dimension of accuracy is particularly critical to sampling because it speaks to whether the sampling frame is an accurate reflection of the population in question, and whether all members of the population have a known, non-zero probability of selection. The QAP includes a number of ideal-world, real-world, and required indicators relating to accuracy, and indicators have been grouped by achievability. Table 1 includes eight indicators designated as required, and all targets were fully met.[1]

[1] The eighth indicator – regarding net/planned sample size – was assessed in Kantar Public's QAR as unmet because two interviews had to be deleted in France in the quality control phase. Reduction of France's sample size from 1200 to 1198 is not significant enough to impact statistical conclusions. Therefore, because there is no true impact on data quality, we have assessed this as 'Target met'.

Table 2. Accuracy Indicators in Sampling Frame Development (Real-World)

Indicator	Target	UM Assessment
RW: 100% of the population covered by the reference statistics	100%	Target met
RW: Specified information on stratification variables included in the register	100%	Target met
RW: Percentage of countries where sampling frame covers at least 95% of population	100%	Target met
RW: Reference statistics used for stratification updated within a year preceding fieldwork	100%	Target not met
RW: Percentage of countries where register updated within a year preceding fieldwork	100%	Target not met
RW: Percentage of countries where specified information on degree of urbanisation is using a common set of categories	100%	Target not met
RW: Percentage of countries where distributions across stratification categories of the net sample approximate the distributions of the universe (deviations in the proportional size of each of the strata between the two should not exceed 1 percentage point)	100%	Target not met

For the seven indicators (Table 2) designated as ‘real-world’, specified targets were achieved for the first two indicators, concerning the extent to which the population was fully covered by reference statistics and for which stratification data were included in the registers (where applicable), In Kantar Public’s QAR, the indicator referring to the percentage of the population to be covered by the sampling frame was self-assessed as unmet because the target was not achieved in Turkey, where of 8% of the population was excluded due to the crisis in neighbouring Syria. This indicator is generally relevant to quality because of the impact that systematic exclusion of certain populations can have on conclusions drawn both within and across countries in analyses. However, in the case of Turkey, the omission of this particular population was known, and reference to the data in Turkey should include the caveat about the elements in the population not represented by this sample. In this context, we consider there to be no impact on overall quality of the data itself, and assess this target as met.

The following three indicators reference the availability of up-to-date reference statistics, population registers, and comparable definitions of urbanisation. Here, targets were not achieved. However, as discussed in Section 4.1, in a review of best practices, here the issues of quality and documentation are more critical (Hubbard et al., 2016), As with many facets of 3MC survey research, use of the same procedures across countries is not necessary for optimizing comparability. The limitations of these indicators to assess quality are discussed in Section 2.5.

The final indicator in Table 2 refers to the net sample. In just five countries, the deviation between the stratification categories of the net sample was less than one percentage point. Post-stratification weights adjust for the differences, but a measure of precision is lost. We discuss this deviation further in Section 3.3 in the assessment of sample composition output.

Table 3. Accuracy Indicators in Sampling Frame Development (Ideal-World)

Indicator	**Target**	**UM Assessment**
IW: Percentage of countries using a high quality register	100%	Target not met
IW: In countries using pre-selected sampling frame, percentage of sampling frame units that refer to non-existent or non-eligible addresses	0%	Target not met
IW: In countries using enumeration, percentage of sampling frame units that refer to non-existent or non-eligible addresses	0%	Target not met
IW: Percentage of enumeration checks reveal deviations from country specific plan	0%	Target not met
IW: Percentage of register entries with a wrong or non-working telephone number	0%	Target not met
IW: Percentage of sampling frame units with incomplete contact information and not otherwise contacted	0%	Target not met

Compliance was not met with regards to achieving any of the targets for ideal-world indicators, which is apropos with the designation of ‘ideal-world (Table 3), The first indicator regarding the use of registers is relevant to quality because the objective is to use a probability sample that is optimal, cost-efficient, and representative of the target population. Additionally, registers facilitate centralised sample selection and management for standardisation of the respondent selection process in a 3MC survey. However, many countries either do not have population registers, or do not have up-to-date or fully comprehensive registers and alternate methods may need to be considered (see Sections 4.1 and 6.1), Considered in isolation, whether or not the target for this indicator was achieved does not provide an indication of data quality.

The following two indicators refer to the extent to which non-existent or non-eligible addresses appear on the register and enumeration sampling frame, resulting in overcoverage. Overcoverage in this case is not necessarily an indicator of data quality per se; indeed, registers are not updated at the rate that may be required to meet the needs of survey organisations and may include some overcoverage. In survey research, overcoverage is primarily a measure of cost in that additional interviewer effort is spent attempting to contact an invalid or ineligible address. However, rates of overcoverage are generally quite low. Indeed, while the targets for these two indicators were not met, rates of overcoverage across all countries were low, ranging from below 4% among EU countries to below 6% for EU candidate countries.

The fourth indicator refers to deviations from country-specific enumeration plans. In post-enumeration verifications ahead of the field period, very few issues were detected, and were rectified as discussed in the QAP, resulting in minimal impact on data quality.

The final two indicators refer only to Sweden, where the telephone was used as a mode of pre-contact. Compliance with the fifth indicator was not achieved due to about 10% of register entries in Sweden having a non-working or incorrect telephone number. Accuracy in telephone registries suffers due to population mobility, and more important to the issue of data quality is that the lack of a telephone number was not subsequently followed by other means of contact with a sampled household or individual. This is captured in the sixth indicator, where compliance was not achieved due to about 6% of sampling frame units unable to be contacted by telephone and not contacted by any other means. The subsequent impact on data quality is seen in the nonresponse statistics reported in Section 3, and further discussed in Section 4.3.

Additional quality indicators addressing coherence and comparability, accessibility, and timeliness and punctuality for sampling frame development are shown in Appendix 1, Table A1. Indicators 1 and 2, relating to coherence and comparability, refer to the comparability of stratification categories across countries. Indicator 1 also appears as a real-world indicator of accuracy, and its limitations are included in the discussion of Table 2 above. Indicator 2 also refers to the issue of a common set of variables used across countries, and a similar statement about the resultant impact on quality applies. That is, use of identical categories across all countries does not necessarily guarantee comparability, and key is deliberation in selection and comprehensive documentation. Indicators 3 to 7 consider accessibility, and speak primarily to documentation of processes, which are of importance to data users. Compliance was achieved for all specified indicators.

Indicators 8 to 13 assessed punctuality in terms of the delivery and approval dates of sampling plans, delivery of the gross sample, training of enumerators, and completion of enumeration and its quality control procedures. Punctuality is an important indicator for data quality in the context of the sequential nature of certain tasks. For example, it is important for sample plans to be agreed upon in advance of data collection, allowing for sufficient time for training materials to be developed, and so on. Punctuality is also important in the context of a 3MC survey, where timing of research activities should be comparable wherever possible so as not to introduce potential sources of measurement error. The available data regarding delivery and approval of sampling plans indicated failure to deliver sampling plans and gain approval prior to deadlines. However, the extent to which any particular deadline is met may be less meaningful in terms of potential impact on data quality. Therefore, only Indicators 8 and 9, related to the sequential nature of tasks, have the potential to impact data quality. As compliance with these two indicators was achieved, we conclude that overall compliance was met with regard to punctuality in sampling frame development.

2.2 Questionnaire development and advance translation, cognitive testing, and translation

In this section, we assess the processes of questionnaire development according to the QAP which include indicators related to development of new questions for the 4th EQLS as well as advance translation, cognitive interviewing, and translation.

Table 4. Accuracy Indicators in the Questionnaire Development Process

Indicator	**Target**	**UM Assessment**
Questionnaire development		
RW: Percentage of questionnaire items in the final source questionnaire that meet international methodological standards of question design (such as outlined in Saris & Gallhofer (2007)	100%	Target not met
Advance translation		
RQ: Percentage of the questions where substantive ambiguities are spotted, for which elaborate documentation of the consideration for translation is provided	100%	Target met
RQ: Percentage of questionnaire items where substantive ambiguities are spotted for which either the source questionnaire is adjusted or a translation instruction is drafted	100%	Target met
Cognitive testing		
RQ: Number of questions for which 'major' issues were detected that were kept	0	Target met
Translation		
RQ: Percentage of translators and adjudicators that take part in the training	100%	Target met
RQ: Percentage of countries where translation is carried out by two translators, out of which one is independent from the national fieldwork agency	100%	Target met
RQ: Percentage of countries where reviewing is carried out by two translators, out of which one is independent from the national fieldwork agency, and an adjudicator	100%	Target met
RQ: Percentage of cross-national review sessions, in which adjudicators from each of the countries sharing the particular language participate	100%	Target met
IW: Percentage of questionnaire items that required editing	0%	Target not met

The indicator addressing the quality dimension of accuracy for **questionnaire development**, shown in Table 4, is that 100% of the questionnaire items in the final source questionnaire meet international methodological standards of questionnaire design. The standards outlined in Saris & Gallhofer (2007) are provided as a possible set of criteria for assessing the outcome for this target. However, based on the criteria we developed to assess the final source questionnaire which draws on Saris and Gallhofer (2007) as well as other essential texts on writing good questions (Groves et al., 2009; Sudman and Bradburn, 1982; Saris and Gallhofer, 2014), we identified some issues that are discussed further in Section 3.1. Based on this assessment, we find that this target was not met.

Table 4 also includes two accuracy-related indicators defined for the **advance translation process**. Based on our review of the documentation related to the advance translation process, there is evidence that thorough commenting to inform translation was generated or the source questionnaire adjusted for all questions for which substantive ambiguities were identified in advance translation. Both targets were therefore achieved.

One indicator addresses accuracy for **cognitive testing**. Information in Kantar Public's Technical and Fieldwork Report clearly outlines the questions for which issues were detected in the cognitive testing and how these issues were addressed in each case indicating that this target was also met.

Most of the indicators related to accuracy with regards to **translation** were required with the exception of one ideal-world indicator. The required indicators are related to participation in training, participation in review sessions, and the number of translators/adjudicators involved in the translation and reviewing process. As the outcomes for these indicators can be measured in terms of people, they can be assessed relatively clearly and objectively. Based on the evidence provided by Kantar, the targets for these required accuracy-related indicators for translation were all met. The ideal-world indicator concerns the percentage of questionnaire items requiring editing. The target for this indicator was 0% however, Kantar Public reports that 3-4% of questionnaire items required editing so this ideal target was not met.

Additional quality indicators addressing relevance and timeliness, accessibility and punctuality for questionnaire development are in Appendix 1, Table A2. One indicator (14) was defined for questionnaire development addressing the quality dimension relevance & timeliness. Evidence provided by Eurofound demonstrates that Eurofound's stakeholders and Advisory Committee were consulted on the development of the questionnaire through expert meetings, a pilot module, and other consultations, thus meeting the specified target.

Indicators 15-16 address the quality dimension of accessibility in advance translation. Based on our review of the available documentation, we conclude that these targets were not fully met because there were no specific instructions that we could find in the documentation provided to us for the assessment and there were several other gaps in the available documentation as we discuss in Section 4.2. Indicator 17, defined for cognitive testing, concerns the questionnaire items for which systematic documentation is provided about the extent to which answers in the cognitive interviews correspond with the concepts that are intended to be captured by the questions. According to Kantar Public, this target was met completely, with all new questionnaire items tested and documented in the Cognitive Report together with a score for each question tested.[2] The evidence provided by Kantar Public that this target was met is systematic in that a score is provided for each item/respondent. However, the process for assigning scores and the extent that the scoring system addresses respondent comprehension of the questions as intended, as specified in the indicator, is not clear, nor is this information available in the materials we have received related to the cognitive testing. We therefore find that this target was not met. We provide recommendations related to this issue in Section 6.2. We find that targets related to the remaining accessibility translation indicators (18-21) were met, based on the selection of translations we reviewed in detail for this assessment.

[2] Score out of 1-5: 1 = very difficult; 5 = very easy.

Targets associated with punctuality-related indicators (22-29) concerning questionnaire development, advance translation, cognitive testing, and translation were all met with the exception of the cross-country review (28) and delivery of the final translated questionnaires (29), which were delivered 5 and 6 days late respectively, a delay with no or negligible impact on data quality.

2.3 Fieldwork (Implementation, monitoring, contact procedures, nonresponse, and paradata)

The survey lifecycle phase of fieldwork implementation and monitoring encompasses a wide range of processes, including development of comprehensive interviewer tools, interviewer recruitment, training, and monitoring, and interview verification. Collecting comparable data in the context of 3MC surveys is a highly complex task, and social, political, geographic, and economic heterogeneity can manifest itself through a wide variety of dimensions that can impact data collection efforts. The QAP indicators regarding fieldwork processes set targets that, if met, promote the goal of data quality and cross-national comparability in the 4th EQLS.

Table 5. Accuracy Indicators in Fieldwork Implementation and Monitoring

Indicator	Target	UM Assessment
RQ: Percentage of national fieldwork managers attending the fieldwork manager instruction meeting	100%	Target met
RQ: Percentage of interviewers that take part in training	100%	Target met
RQ: Training materials cover instructions on CAPI program/questionnaire	Yes	Target met
RQ: Training materials cover instructions on consistency checks	Yes	Target met
RQ: Training covers all relevant materials	Yes	Target met
RQ: Training materials cover strategies for convincing reluctant respondents	Yes	Target met
RQ: Training materials cover guidelines on contacting process	Yes	Target met
RQ: Percentage of gross sample entries that are discarded before the net sample is realised, for which a final outcome has been realised (i.e. no cases to be lost)	100%	Target met
RQ: Percentage of sample entries to which a final status of 'noncontact' was assigned that were not visited at least four times at different times and on weekdays and weekends	0%	Target not met
RW: Percentage of countries using a common integrated CAPI and sampling management system	100%	Target not met
RW: Percentage of issues identified based on information in weekly	100%	Target met

monitoring data for which a solution is provided		
RW: Percentage of countries where at least 10% of interviews are checked and for which the first recontact attempt was made within a week after the interview was carried out	100%	Target not met

The QAP included twelve indicators relating to accuracy relative to fieldwork. The first two indicators in Table 5 refer to participation in a fieldwork manager meeting and in interviewer trainings, and compliance here is achieved. The following five indicators refer to the extent to which the interviewer training materials were comprehensive, which in turn promotes consistent behaviour among interviews, leading to increased accuracy. All targets for these indicators were achieved as well.

The following two indicators concerned assignment of final sample disposition codes. The first indicator, referring specifically to the percentage of gross sample entries discarded before the net sample is realised, for which a final outcome has been realised, was achieved. The QAR notes that only one address did not have a final outcome, which clearly has no impact on data quality. The other indicator is a measure of the extent to which final outcome codes were assigned without contact procedures being followed as prescribed. A relatively low percentage of these cases are reported in the QAR and the Data Editing and Cleaning Report and as evidenced by our independent analyses of these data. However, some countries experienced a greater number of specific types of deviations from the protocol. For example, in Austria and France, there were 183 and 242 cases, respectively, where a final response code was assigned after three or fewer contact attempts, in a departure from the specified minimum of four contact attempts.[3] The Data Editing and Cleaning Report notes that such deviations generally occurred towards the end of the fieldwork, which might suggest that these addresses were closed out once the target sample size was reached. Nonresponse bias occurs when non-respondents differ systematically from respondents. When interviewers do not follow the specified protocol with regard to hard-to-reach respondents, the risk of nonresponse bias will increase. Section 6.3 includes a discussion of both potential nonresponse bias analyses as well as methods to limit this form of error in future surveys.

The tenth indicator refers to the extent to which a common CAPI instrument and sample management system was used across all countries. This target was reported as unmet because of alternate hardware usage in two countries. Subsequent discussions with Eurofound indicate that the use of alternate hardware did not impact data quality. More important is an assessment of the technical development process, for which documentation is lacking in the 4th EQLS. We discuss further best practices and recommendations in Sections 4.3 and 6.3.

The final two indicators refer to data monitoring and verification. The first indicator speaks to the need for resolution on all quality issues, and the QAR reports full compliance with this indicator, but no documentation is available. The final indicator refers to the percentage of interviews verified within a set timeframe. The evidence provided in the QAR indicates that protocol was followed for about 90% of interviews, meaning that compliance was not fully achieved. Specific data regarding interview

[3] These figures were calculated using variables RESPONSECODE, y16_country, and Y16_CS_contactattempts.

verification is not available, and Section 6.3 discusses how processes and documentation might be improved in relation to these final indicators.

Additional quality indicators addressing accessibility and punctuality for fieldwork are shown in Appendix 1, Table A3. Issues of accessibility with regard to fieldwork are important from the perspective of documentation for data users. Indicators 30 and 31 refer to the availability and provision of the recruitment materials used for the 4th EQLS, and there is compliance with both. The third indicator (32) refers to the percentage of countries included in weekly monitoring data. The target as evidenced in the QAR is unmet, although details indicate that this was a rare occurrence and that there was no impact on data quality. The final indicator (33) regarding accessibility refers to delivery of the methodological and fieldwork report, and here compliance was achieved.

The QAP included eight indicators relating to punctuality relative to fieldwork (34 – 41), As discussed earlier, punctuality as a quality dimension is primarily relevant in the context of task sequencing and cross-national comparability. Indicators 34 and 35 speak to this and are important indicators of quality in that it is crucial for both fieldwork manager meetings and interviewer training to occur *before the start of the fieldwork*, as the target specified. Indicators 36 and 37, which details specific dates for interviewing material and CAPI programming delivery, were not met. However, the important aspect of both of these indicators is that materials are complete in advance of interviewer training and commencement of fieldwork. As the materials were indeed delivered in time to meet these obligations, the delay in delivery has no or negligible impact on data quality. Indicators 38 and 39 refer to the timeliness of the delivery of monitoring data and subsequent communication between Kantar Public and Eurofound. Full compliance was not achieved, but there were only a few very minor delays, most of which occurred because of holidays. Again, this should have no or negligible effect on data quality. Indicator 40 refers to the number of days which fieldwork extended beyond the initial fieldwork end date. The impact on data quality cannot be ascertained simply by assessing the number of calendar days. We discuss this further Section 6.3. The final indicator (41) refers to the delivery date of the Technical and Fieldwork Report. The target was not achieved, but this was the result of the extended fieldwork period. The report was eventually delivered and therefore there is no impact on data quality.

2.4 Weighting

The distribution of groups of observations in a survey dataset may differ from the distribution in the survey population due to the quality of the sampling frame, the sample design, and patterns of unit nonresponse. Weighting can help correct for these differences as well as reduce the sampling bias of the estimates and compensate for non-coverage and unit nonresponse. The original tender for the 4th EQLS specified three types of weights to be estimated by the contractor – design weights, post-stratification weights, and cross-national weights – with external sources specified for post-stratification weights.

Table 6. Accuracy Indicators for Weighting

Indicator	**Target**	**UM Assessment**
RQ: Percentage of countries where weighting strategy integrates all available information on elements foreseen to be included in the weighting procedure, given the sampling plan.	100%	Target met
RQ: Percentage of countries where design weight is specified in accordance with sampling design	100%	Target met
RQ: Percentage of countries where the post-stratification weight takes all variables, requested in Terms of Reference / agreed, into account.	100%	Target met
RQ: Weight trimming follows the weighting strategy and is fully documented and replicable	Yes	Target met
IW: Percentage of countries where a common set of variables with common categories are used for weighting	100%	Target not met
IW: Percentage of countries where the weights are based on up-to-date official population statistics collected within two years preceding fieldwork	100%	Target unknown

The accuracy indicators for weighting include four that were required and two that were categorised as ideal-world, as shown in Table 6. The first accuracy indicator relates to the percentage of countries for which the weighting strategy integrates all available information on the elements to be included in the weighting procedure, as outlined in the sampling plan. Based on our review, the syntax and calibration files appear to be comprehensive of all available information for all countries, thus meeting the target. The next two required accuracy indicators include the percentage of countries for which the design weight was specified in accordance with the sampling plan and for which the post-stratification weight takes into account all variables requested in the Terms of Reference in the initial tender. Our assessment of the syntax and final dataset files indicate that the design and post-stratification weights were calculated accurately and according to the specifications agreed with Eurofound for all countries, meeting the targets for both of these indicators. Trimmed weights were included in the dataset and documented in the Sampling, Enumeration, and Weighting report, meeting the target defined in the fourth required accuracy indicator.

The first-ideal world indicator is that a common set of variables with common categories were used for weighting in all countries. According to Kantar, this was possible for each of the EU28 countries but not in all of the five EU candidate countries due to difficulty in obtaining household size information. Thus, the target for this ideal world indicator was not met. However, as discussed in Section 4.4, differences among countries in the type and quality of external data available for use in post-stratification adjustments is not uncommon. The key is that weighting adjustments are made based on the best available data in each country.

The second ideal-world indicator concerns the quality of the official population statistics upon which the weights are based for each country; specifically, that the official statistics were collected within two years of the fieldwork. According to Kantar, this target was met but we are unable to corroborate this. As far as we can tell, the details regarding the specific source (e.g., URL, database, etc.) and date extracted for the control data used was not provided in the available documentation.[4]

Additional quality indicators addressing coherence and comparability, accessibility and punctuality for weighting are shown in Appendix 1, Table A4. The indicator addressing coherence and comparability (42) relates to the target that the weighting strategy includes references to the academic literature demonstrating that the selection of weighting variables and procedures adheres to common practice for international surveys. The weighting strategy was revised for the 4th EQLS based on an expert review and recommendations (Vila & Cervera, 2014) following the 3rd EQLS. Overall, we find that the weighting strategy for the 4th EQLS reflects current best practice as discussed in the academic literature and practiced by other major cross-national surveys. See Section 4.4 for further discussion and Section 6.4 for recommendations.

Transparency and clear documentation is particularly important when it comes to weighting. We find that all of the accessibility targets (43 – 51) were met. None of the punctuality targets (52 – 55) were met but the delays of approximately two weeks are minor and would not impact data quality. Fortunately, delays at the weighting stage do not affect the bulk of the survey lifecycle, which precedes it.

2.5 Quality assurance framework assessment

Eurofound has made important strides in advancing survey quality in the EQLS through the development and implementation of its QAP. As our assessment of compliance detailed above demonstrates, most of the indicators are achievable, while the distinction between required, real-world, and ideal-world targets recognises that certain targets are more difficult to achieve than others.

Available documentation regarding the development of the QAP suggests that it was intended to be a quantitative guide for the tenderer in self-assessment of achieving quality in the survey process, as well as a framework for the external evaluation of the tenderer's achieved compliance with the specific quality indicators. We find the QAP, in general, to be a useful tool for monitoring survey process quality as initially envisioned. However, as the external assessors, we find that the QAP could be complemented with additional information. We discuss this further in Section 6.7.

As we assessed compliance with the targets for the quality indicators specified in the QAP, we came across a number of challenges. Specific issues we observed include:

- Single indicators combined in more than one target (e.g., t*ranslation materials are constructed using input from the cognitive test and advance translation, are provided to the translators, and are made publicly available*),
- The outcome is not measurable or criteria for meeting the target is not clearly specified (e.g., *percentage of questionnaire items in the final source questionnaire that meet international methodological standards of question design*),

[4] Eurofound has subsequently updated the Sample evaluation, enumeration and weighting report.

- Related to the item above, some indicators included language that could be considered subjective or open to interpretation (e.g., *systematic documentation be provided* and *clear instructions be provided),*
- Some indicators addressed more than one quality dimension (e.g., *percentage of countries where the specified information on degree of urbanisation uses a common set of categories* was designated as both accuracy and coherence and comparability),
- For some indicators, there is inconsistency between the unit of measurement in the target and the available data required to assess the quality indicator (e.g., evidence provided by Kantar Public for achieving the indicator *Percentage of countries covered in weekly monitoring data (in accordance with template)* refers to the number of weeks in which reports were complete, rather than the number of countries),
- The link between target achievability (i.e., ideal/real-world/required) or whether or the extent to which targets were achieved and the impact on data quality is not always clear.

Recommendations in Section 6.7 include a discussion of how to revise the QAP in order to maximise its effectiveness in future surveys.

3. Assessment of survey outputs

We turn now from an assessment focused primarily on the processes used to collect the 4th EQLS data to a review of several different types of survey outputs, including the English-language source questionnaire, nonresponse statistics, sample composition statistics, and several other types of statistical output. Not only does this exercise permit an evaluation of the quality of the data, but is another method used to further assess the processes used to collect the data.

3.1 Source questionnaire assessment

About half of the survey questions in the 4th EQLS had been fielded in previous wave(s), while the remainder were developed specifically for inclusion in this most recent survey. In this assessment, we elected to focus exclusively on these new items in an expert review of the source questionnaire as they were a direct output of the questionnaire development process. To assess the quality of the final source questionnaire, we draw upon the comprehensive set of guidelines for writing good questions developed by Sudman and Bradburn (1982) and summarised and updated by Groves et al. (2009), as well as principles regarding questionnaire design drawn from Saris and Gallhofer (2014), to develop criteria to guide development of survey questions, particularly as they relate to 3MC surveys such as the EQLS.

Listed in Appendix 2, these guidelines are organised according to question type: 1) nonsensitive questions about behaviour; 2) sensitive questions about behaviour; and 3) attitude questions. As Groves et al. (2009) explain, it is useful to distinguish between these question types because they tend to raise different issues. For example, sensitive questions are more prone to deliberate misreporting and may need to be handled differently to encourage honest reporting. Attitude questions, on the other hand, are more likely to involve different types of response scales, which bring up particular issues. Although these criteria are mapped to specific question types, we note that the distinctions are not mutually exclusive but provide a general framework with which to assess question quality. New questions in the 4th EQLS focused primarily on the quality of public services, and included items measuring both behaviour (e.g., *Have you or someone else in your household used GP, family doctor, or health care services in the last 12 months?*) and attitudes (e.g., *How satisfied were you with the quality of the (GP) facilities?*), with very few items measuring attitudes considered to be sensitive.

We first note several areas where decisions made about questionnaire design can have methodological implications for comparability:

- Measurement error can occur when the order of questionnaire items changes from one survey to another, as the context in which an item occurs can affect response (Tourangeau et al., 2000), Many of the items new to the 4th EQLS were included near the end of the questionnaire, meaning that items asked in previous waves would not be subject to potential order effects introduced by these new items. However, some items (e.g., Q7b) were inserted into the middle of existing batteries of items, which could potentially introduce order effects and impact comparability between waves. It is typically considered best to avoid changes to question order as much as possible
- Notation alongside a number of items in the questionnaire indicated changes in wording from wave to wave (e.g., Q7a), as well as changes to response options (e.g., Q9), While changes to

wording may contribute to improved comprehension and measurement among respondents, such changes can also introduce comparison error across waves in a cross-sectional survey. Such changes and associated justification should be thoroughly documented for data users.

- In order for interviewers to achieve standardisation across all interviews, quantitative survey questions need to be fully scripted, which was not the case for some batteries of survey items (e.g., Q91, Q92),
- Proxy reports of behaviour and experiences are vulnerable to recall error, and whether such reports are appropriate for analysis should be considered (e.g., Q60),

Our item-by-item assessment also uncovered a number of items for which a survey methodologist trained in 3MC questionnaire design may have suggested revisions, including but not limited to the following observations:

- Some items contain complex assertions that may be cognitively difficult for respondents (e.g., meaning of *important problems* in Q7f)
- Definitions for concepts are not fully comprehensive (e.g., *corruption* in Q63b)
- Some concepts are not fully defined, but it is difficult to advise on revisions without having a clear idea of the analytical objectives, as discussed further in Section 4.2 (e.g., Q30e)
- Some items introduce potential contradictions in a question (e.g., in Q38a, asking about *relatives* only may be sufficient, since household members, who are likely *family members*, are excluded)
- Some items are double-barreled – that is, combine more than one concept (e.g., Q62a-d)
- Some items had ambiguous time frames (e.g., Q84), while others were lacking a specific reference to time frame (e.g., Q7g, Q46)
- Some items which follow screener questions (e.g., Q44 and Q42/43, respectively), lack specificity. In this example, if a respondent reports multiple types of care responsibilities in Q42/43, then it is not clear which type of care a response to Q44 assumes. A similar issue occurs in other series (e.g., Q73 if both respondent and someone else used a service in Q68)
- Some items include hidden assumptions (e.g., Q55a/Q55b, where the issue would be resolved by modifying ‘when’ to ‘if’)
- Middle categories are not always advisable when for specific questions the middle category may be a default ‘don’t know’ (e.g., Q67)

We also note that there were certain items whose quality was difficult to assess because the relevant research objective aims and analysis intentions were not clear (e.g., Q41), Recommendations in Section 6.2 discuss how in the future, the development of new items, and their placement in the questionnaire, would benefit from an expert assessment.

3.2 Nonresponse statistics (Response, refusal, cooperation, and contact rates)

This section examines nonresponse statistics from the 4th EQLS including overall nonresponse rates, and rates of refusal, contact, and cooperation. We begin by comparing nonresponse statistics from the 4th EQLS and the 3rd EQLS. We then compare nonresponse statistics between the last two rounds of the ESS (Round 7 and 8) and compare changes in nonresponse between rounds obtained by the ESS and the EQLS, followed by a comparison of nonresponse statistics from the 4th EQLS to the most recent ESS. We use ESS Round 8 data to the extent possible and compare with data from Round 7 (European Social

Survey, 2016d) for countries for which Round 8 data (European Social Survey, 2017b) are not yet available.

To compare nonresponse, refusal, cooperation, and contact rates, it is important that these rates are calculated in exactly the same way. To compare nonresponse, refusal, cooperation, and contact rates between rounds of the EQLS and between the EQLS and the ESS, we apply the standard definitions and formulas of the American Association for Public Opinion Research (AAPOR) for the response rate (RR1), refusal rate (REF1), cooperation rate (COOP1), and contact rate (CON1), The formulas used and the outcome codes for the 3rd and 4th EQLS and the 7th and 8th Rounds of the ESS and the corresponding AAPOR nonresponse categories are provided in Appendix 3. The tables below present the differences in the nonresponse statistics being compared. Appendix 3 provides tables (Tables A6 and A7) with the response, refusal, contact, and cooperation rates for the 3rd and 4th EQLS and the ESS Round 7 and Round 8.

Table 7 shows the differences in response, refusal, contact, and cooperation rates, response rates between the 3rd and 4th EQLS.[5] We see that response rates went up or stayed the same in 12 countries out of 31 between the 3rd and 4th EQLS, as indicated in green. However, response rates went down in 20 countries, as indicated in red, 11 of which saw a decline of 10% or more. Refusal rates went down or held constant in 15 or nearly half of the countries. The decline was considerable in several countries including Montenegro, Serbia, Spain, Turkey, and Hungary. But refusal rates rose in the balance of countries, 11 of which saw increases of 10% or more.

Table 7. Difference in Response, Refusal, Contact, and Cooperation Rates Between the 4th and 3rd EQLS

Country	**RR1 Difference**	**Country**	**REF1 Difference**	**Country**	**CON1 Difference**	**Country**	**COOP1 Difference**
Montenegro	0.26	Montenegro	-0.39	Portugal	0.28	Montenegro	0.35
Serbia	0.23	Serbia	-0.22	Latvia	0.15	Serbia	0.22
Spain	0.16	Spain	-0.19	Bulgaria	0.14	Spain	0.22
Turkey	0.16	Turkey	-0.19	UK	0.14	Turkey	0.17
Czech Rep	0.14	Hungary	-0.18	Denmark	0.10	Czech Rep	0.16
Portugal	0.14	Czech Rep	-0.15	Austria	0.10	Hungary	0.15
Hungary	0.13	Luxembourg	-0.08	Ireland	0.08	Luxembourg	0.09
Luxembourg	0.07	Croatia	-0.03	Germany	0.06	Croatia	0.02
Denmark	0.05	Slovenia	-0.03	Romania	0.05	Denmark	0.01
UK	0.05	France	-0.03	France	0.04	UK	-0.01

[5] Albania and Kosovo are not included as the 3rd EQLS was not carried out in Albania and the 4th EQLS as not carried out in Kosovo.

Croatia	0.04	Sweden	-0.02	Croatia	0.04	France	-0.03
France	0.00	Lithuania	-0.02	Serbia	0.03	Portugal	-0.03
Latvia	-0.01	Denmark	-0.01	Estonia	0.03	Netherlands	-0.04
Netherlands	-0.02	Belgium	0.00	Netherlands	0.03	Slovenia	-0.04
Bulgaria	-0.02	Greece	0.00	Slovakia	0.02	Lithuania	-0.06
Slovenia	-0.04	Slovakia	0.01	Luxembourg	0.02	Finland	-0.06
Romania	-0.04	Netherlands	0.02	Turkey	0.01	Slovakia	-0.08
Finland	-0.05	Finland	0.03	Czech Rep	0.01	Belgium	-0.08
Ireland	-0.05	FYROM Macedonia	0.06	Finland	-0.01	Romania	-0.10
Slovakia	-0.06	Romania	0.06	Poland	-0.01	FYROM Macedonia	-0.12
Lithuania	-0.07	UK	0.09	Spain	-0.01	Bulgaria	-0.14
Belgium	-0.10	Italy	0.10	Slovenia	-0.01	Ireland	-0.14
Estonia	-0.12	Ireland	0.10	Italy	-0.01	Greece	-0.14
Italy	-0.13	Malta	0.12	FYROM Macedonia	-0.03	Italy	-0.14
FYROM Macedonia	-0.14	Estonia	0.12	Hungary	-0.03	Latvia	-0.15
Austria	-0.16	Portugal	0.13	Malta	-0.04	Malta	-0.18
Greece	-0.18	Cyprus	0.13	Lithuania	-0.04	Estonia	-0.18
Malta	-0.18	Bulgaria	0.14	Belgium	-0.06	Cyprus	-0.22
Germany	-0.23	Latvia	0.15	Cyprus	-0.08	Austria	-0.26
Cyprus	-0.26	Germany	0.20	Montenegro	-0.10	Germany	-0.27
Poland	-0.28	Austria	0.21	Greece	-0.15	Sweden	-0.27
Sweden	-0.30	Poland	0.26	Sweden	-0.26	Poland	-0.31

Contact rates increased in 20 countries and decreased by 1% or less in an additional five countries. Sweden saw the greatest decrease in the rate of contact by 26%, which is most likely due to initial recruitment being carried out by telephone. We address the issue of the mode of initial contact and provide recommendations in Section 6.3. Cooperation rates largely fell between the 3rd and 4th EQLS. This was the case in 22 countries and the decline was sizable in a number of countries with the cooperation rate falling by 15% or more in eight of those.

Overall, the trend in response statistics between the 3rd and 4th EQLS paints a mixed picture. Response rates increased in 12 countries and refusal rates went down or held constant in approximately half of the study countries. This may be attributable to improved field procedures and increased contact rates, which largely increased or stayed nearly constant (within 1%) in 25 countries. Nonetheless, cooperation rates largely suffered in the 4th EQLS. These results underscore the importance of continued focus on response rates for the EQLS and the implementation of efforts to better understand the potential effects of nonresponse bias and to further improve response rates in future waves, as discussed further in Section 6.3.

Comparing response rates between Round 7 and 8 of the ESS, as shown in Table 8, we see that response rates went up or stayed the same in 11 out of 16 countries. Refusal rates fell or held constant in seven countries but increased in ten. The contact rate increased or stayed the same in 9 countries and went down in seven. However, the declines in the contact rate were all less than 5%. Cooperation rates increased in six countries, remained constant in four and decreased in just six countries.

The trend in response statistics between ESS Round 7 and 8 also presents a mixed but perhaps slightly more positive picture than the EQLS. Response rates fell in fewer countries in the ESS than for the EQLS and where response rates did go down, the decline was not as sharp. Refusal rates also increased in 10 countries out of 16 but by less than seen for the EQLS. Gains in the contact rate for the ESS are evident in fewer countries than for the EQLS but the cooperation rate for the ESS increased or held constant for a higher proportion of countries than did for the EQLS. Overall, we see less extreme variation in each type of response statistic for the ESS. It should be noted, however, that the EQLS covers a much larger and more diverse set of countries including a number of countries with less established survey research tradition.

Table 8. Difference in Response, Refusal, Contact, and Cooperation Rates Between ESS Round 8 and 7

Country	RR1 Difference	Country	REF1 Difference	Country	CON1 Difference	Country	COOP1 Difference
Estonia	0.08	Ireland	-0.06	Israel	0.07	Estonia	0.08
Poland	0.04	Estonia	-0.03	Austria	0.05	Ireland	0.06
Slovenia	0.03	Poland	-0.03	Slovenia	0.03	Poland	0.05
France	0.03	Norway	-0.03	Estonia	0.03	Belgium	0.03
Ireland	0.03	France	-0.02	France	0.01	France	0.03
Austria	0.01	Belgium	-0.02	UK	0.01	Slovenia	0.02
Belgium	0.01	Switzerland	0.00	Czech Rep	0.01	Germany	0.00
Czech Rep	0.00	UK	0.01	Switzerland	0.00	Switzerland	0.00
UK	0.00	Czech Rep	0.01	Poland	0.00	Czech Rep	0.00
Israel	0.00	Slovenia	0.01	Norway	-0.01	UK	0.00
Switzerland	0.00	Germany	0.02	Sweden	-0.01	Norway	-0.01
Germany	-0.01	Finland	0.03	Finland	-0.02	Austria	-0.03

Norway	-0.01	Netherlands	0.03	Germany	-0.02	Finland	-0.04
Finland	-0.05	Sweden	0.04	Netherlands	-0.03	Israel	-0.06
Sweden	-0.06	Austria	0.05	Belgium	-0.03	Sweden	-0.06
Netherlands	-0.08	Israel	0.08	Ireland	-0.04	Netherlands	-0.07

Table 9 shows the differences in response, refusal, contact, and cooperation rates between the 4th EQLS and ESS Round 8 data, where available, and from Round 7 for countries for which Round 8 data are not yet available. Direct comparisons are limited to countries where the 4th EQLS was conducted and either ESS Round 8 or ESS Round 7 was carried out. Looking at the differences in response rates, we see that lower response rates were obtained for the EQLS than the ESS in every country except for two. The EQLS also saw higher refusal rates in every country except two, lower contract rates in every country except three, and lower cooperation rates in every country except one compared to the ESS.

Comparing response statistics directly within countries, it is clear that the ESS obtained higher rates of response, contact and cooperation rates in most countries in the most recent round than did the EQLS. There is a large difference in these outcomes even in the case of Belgium and the Netherlands where Kantar Public carried out the fieldwork for both the ESS and the EQLS. That the ESS has tended to achieve higher response rates is interesting given that the EQLS appears to have exerted tighter control over field procedures than the ESS, as discussed in Section 4.3. It would be potentially useful for the EQLS to further explore the details of implementation processes between ESS and EQLS that might help understand the source of these differences.

Table 9. Difference in Response, Refusal, Contact, and Cooperation Rates Between the 4th EQLS and ESS Round 7 or 8

Country	RR1 Difference	Country	REF1 Difference	Country	CON1 Difference	Country	COOP1 Difference
Portugal*	0.06	Czech Rep	-0.03	Germany	0.09	Portugal*	0.09
Hungary*	0.04	Portugal*	-0.01	Poland	0.07	Hungary*	0.00
Spain*	-0.04	Slovenia	0.00	Hungary*	0.06	Czech Rep	-0.03
Slovenia	-0.06	Lithuania*	0.03	Denmark*	-0.05	Slovenia	-0.07
Czech Rep	-0.10	Sweden	0.04	UK	-0.05	Ireland	-0.10
Germany	-0.12	Hungary*	0.05	Austria	-0.05	Spain*	-0.11
UK	-0.12	Austria	0.07	Portugal*	-0.06	UK	-0.12
Ireland	-0.13	Spain*	0.09	Ireland	-0.07	Denmark*	-0.14
Denmark*	-0.15	Ireland	0.09	Belgium	-0.07	Belgium	-0.15
Belgium	-0.16	Denmark*	0.12	Spain*	-0.07	Germany	-0.15
Austria	-0.19	France	0.12	Slovenia	-0.07	Finland	-0.18
France	-0.21	UK	0.13	France	-0.08	Austria	-0.18

Finland	-0.21	Belgium	0.13	Estonia	-0.09	Lithuania*	-0.18
Estonia	-0.23	Finland	0.14	Netherlands	-0.09	Sweden	-0.21
Netherlands	-0.23	Estonia	0.15	Czech Rep	-0.11	Estonia	-0.21
Sweden	-0.25	Netherlands	0.15	Finland	-0.13	Netherlands	-0.22
Poland	-0.29	Germany	0.19	Lithuania*	-0.24	France	-0.22
Lithuania*	-0.30	Poland	0.39	Sweden	-0.28	Poland	-0.39

* Results from ESS Round 7

3.3 Assessment of sociodemographic sample composition

One important aspect of survey data quality is the extent to which the survey sample is representative of the target population. As few data were collected to facilitate a nonresponse bias analysis in the 4th EQLS, we consider instead a comparison of the EQLS data with comparable data stemming from two different sources, data used to calculate the calibration weights from Eurostat, and survey data from the ESS, with sociodemographic variables used in the analysis listed in Table 10. Any emergent patterns in such a comparison between the target and sample populations may be indicative of potential quality control issues in sampling frame development and/or the fieldwork processes.

Table 10. Data Used in Calibration Weight Calculation

Sociodemographic data	**Response categories**
Age by gender	M/F18-29, M/F30-39, M/F40-49, M/F50-59, M/F60-69, M/F70+
Household size	1 person HH, 2 person HH, 3 person HH and 4+ person HH
Employment status	Employed, other

Calibration weights were calculated after data collection using several sociodemographic variables, including age, gender, household size, employment status, and geographic region. While application of calibration weights in analysis is effective at aligning the survey population to the target population in the analysis stage, here we consider further the data used to create these weights—specifically, data pertaining to age and gender—in order to assess how well the sample population is representative of the target population on key sociodemographic data, as a component of our evaluation of both the processes involved in data collection as well as of the overall quality of the data itself.

Table A8 in Appendix 4 compares the proportional age and gender distributions sourced from Eurostat and unweighted data from the 4th EQLS, with percentages indicating the extent to which the survey population deviated from the target population.[6] For example, young males (18 – 29) are

[6] Due to the lack of comprehensive documentation of reference statistics, data from Eurostat to develop the calibration weights was not available for Serbia. Thus, Serbia is not included in Table A8.

underrepresented in the sample population in Austria by about 5.1% (Table A8),[7] However, the differences between the target and sample population are relatively small, with just a few exceptions. Generally, and in line with other surveys, including the ESS, younger people are generally underrepresented, while older people, and especially older females, tend to be overrepresented (Koch, 2016), There are no patterns observed between countries using a register as a sampling frame versus those using an enumeration method, nor are there differences between EU and EU candidate countries. This evidence suggests that potential differences between the EQLS and Eurostat data are due to nonresponse and not indicative of concerns with either the sampling frame or the fieldwork process in the 4th EQLS.

We turn next to data from Round 7 of the ESS and compare several sociodemographic variables to 4th EQLS data. Tables A9 to A12 (Appendix 4) show the results of two sets of analyses, comparing first unweighted and then weighted data.[8] Round 7 of the ESS calculated calibration weights for gender, age, education, and geographical region based on EU Labour Force Survey (ELFS), meaning that the application of calibration weights in a comparative analysis may mask true differences in process and/or output (European Social Survey, 2014), However, a comparison of the weighted data may indicate the comparability of the data used to create calibration weights in the EQLS and ESS.

Table A9, a comparison of the proportional distribution of age and gender using unweighted data, shows little evidence of a pattern of difference between the two surveys, with the majority of differences in cells below 5%. Differences in weighted proportional distributions (Table A10) show even fewer differences. This analysis suggests little difference between the ESS and EQLS in terms of sampling frame development, fieldwork processes, and comparability of data used to calculate calibration weights.

In a comparison of the unweighted proportional distribution of household size (Table A11), we see a number of differences between the two surveys. In the majority of EQLS countries, single-member households are overrepresented in the sample population as compared to the ESS. There is more similarity, however, in terms of employment status, with the majority of countries having less than a 5% difference between the two surveys. Table A12, comparing weighted proportional distributions in household size and employment status, show more consistency between the two datasets, indicating comparability between the sources used to develop the calibration weights.

3.4 Evaluation of other output statistics

In addition to a basic analysis of nonresponse and sample composition, we discuss here other types of output analysis that provide an indication of data quality in an overall assessment, including analysis of

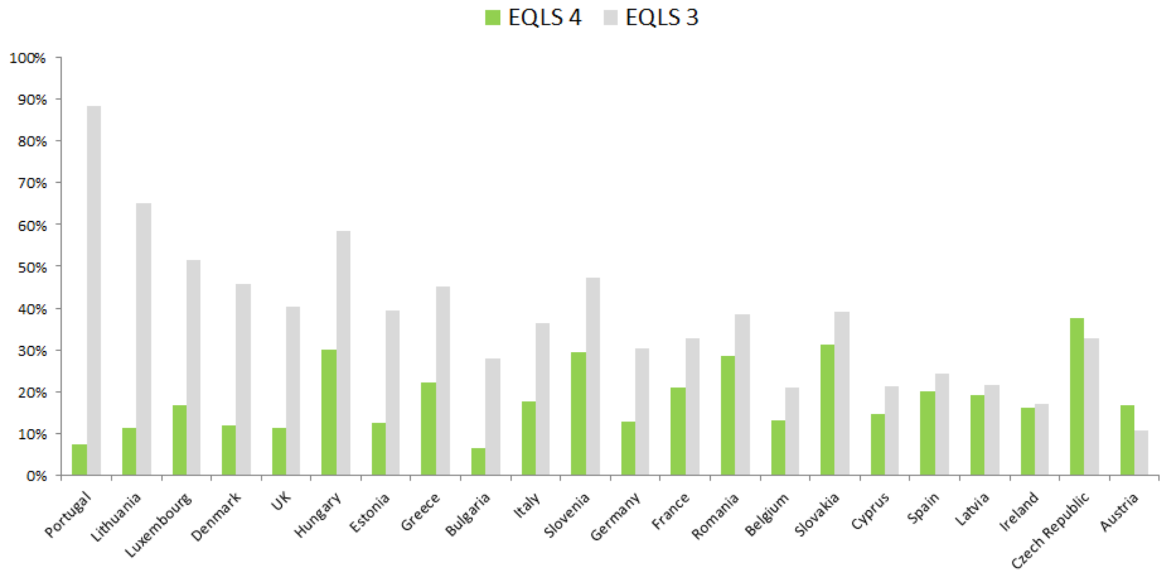
[7] While not shown in these tables, according to the data from Eurostat, males aged 18 – 29 comprise about 9.4% of Austria's population (aged 18+), but only about 4.2% of the 4th EQLS sample was comprised of males in that age range, for a difference of about -5.1%.

[8] The ESS7 – integrated file, Edition 2.1, was used to complete these analyses, as at the time of this assessment, calibration weights were not yet available for ESS Round 8. Note that there is not universal overlap between ESS Round 7 and the 4th EQLS in terms representation of countries; we include in this analysis only those countries which participated in both surveys, in these specific rounds.

the coefficient of variation and design effects at the country level, interviewer effects, item nonresponse and straight-lining analysis, and the design effects and standard errors for specific survey items. Although relevant to our assessment in general, in order to explain our approach to output analysis we specifically note here that analysis of output data serves several purposes: 1) how users of EQLS data might consider issues of quality in specific countries; 2) how Eurofound can use these data to inform the design of future surveys; and 3) how we, as external evaluators, can use these data to assess overall quality of the EQLS.

Analysis of the coefficient of variation, comparing the extent of variability in relation to the mean of the population in the issued sample, is a relative precision indicator. A comparative analysis was conducted by Eurofound prior to this assessment, using data from both the 3rd and 4th EQLS, and for efficiency we cite it here rather than replicate it (Ahrendt, et al., 2017), As Figure 2 shows, overall the coefficient of variation is relatively stable across participating countries in the 4th EQLS, with estimates from all countries lower than 40%, and lower than 30% in the majority of countries, leading to greater comparability. There is also a significant reduction in variability in all but two countries (Czech Republic and Austria) when comparing data from the 4th to the 3rd EQLS. The adjustments made in the 4th EQLS with regards to sample release, which attempted to adjust for nonresponse differences across PSUs, appears successful in yielding a more efficient sample and increased precision, as evidenced in the smaller coefficient of variation.

Figure 2. Coefficient of Variation of Issued Sample[9]

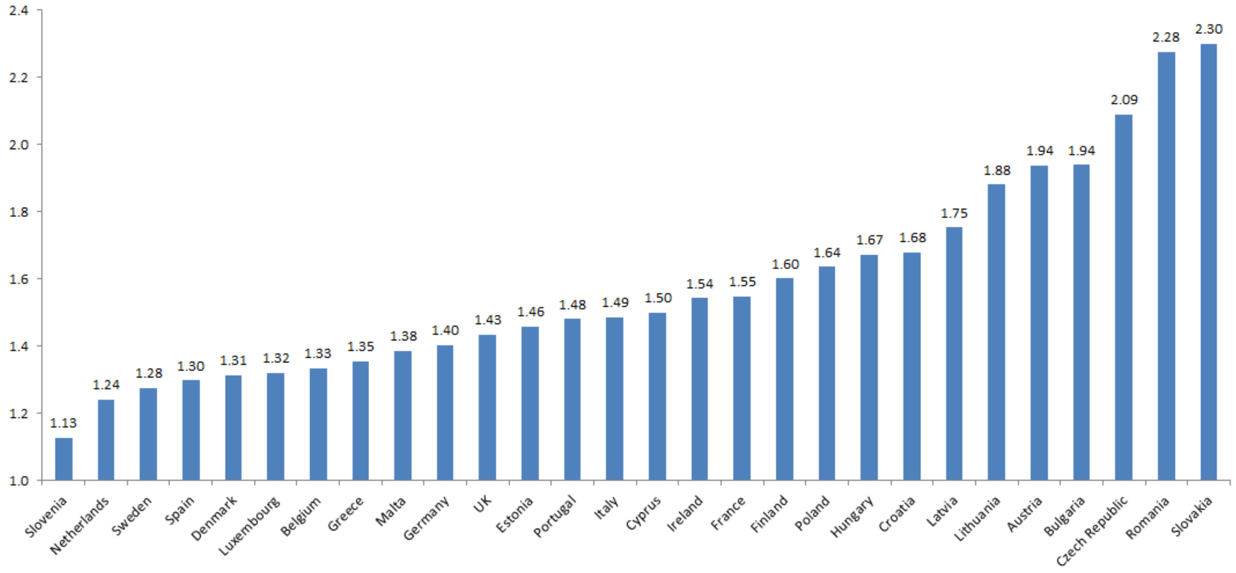


Analysis of design effects at the country level provides an indication of the comparability of efficiency achieved in the sampling process, which is valuable for Eurofound when considering sample design options and implementation. Eurofound calculated design effects at the country level as well, and we again draw on their analysis rather than replicate it ourselves (Ahrendt, et al., 2017), Figure 3 shows a

[9] The coefficient of variation is the ratio of the standard deviation to the mean = stdev/mean.

relatively consistent distribution of design effects across countries. While the Czech Republic, Romania, and Slovakia all have slightly higher design effects, all range between 1 and 2.5. These results do not indicate any issues of data quality, and suggest that the sample design and implementation processes followed in the 4th EQLS are satisfactory.

Figure 3. Design Effects in the 4th EQLS[10]



Previous assessments of the EQLS (Petrakos, et al., 2010; Vila, et al., 2013) conducted an extensive analysis of standard errors and design effects for a wide range of variables for each participating country as well. In surveys where there are just a few key variables of interest, an analysis of these data can inform interpretation (e.g., a survey focused on a specific health outcome, with several key variables relating to prevalence of a condition), However, in an omnibus survey such as the EQLS, such a broad analysis is statistically likely to produce a number of outliers or extreme values, many of which are apt to be spurious and due to statistical chance, rather than due to any issue in the survey process. Evaluators with methods training, who are not experts in the subject area of the survey are less likely to be able to determine when extreme values reflect actual phenomenon versus a spurious result.

Considering the needs of end users and what might be gained by such an analysis in this assessment, we note that modern statistical analysis packages provide users with tools they need to account for complex sample designs, calculate the standard errors, etc. An end-user will use these tools in a focused way, based on a specific analysis plan driven by informed hypotheses and alongside the substantive background knowledge required to best interpret the output, leading to is little utility in cataloguing these a priori. Furthermore, the calculation of standard errors (to use in significance testing) exists for each and

[10] Design effects were calculated based on the Kish's approximate formula: Design effect = $\frac{\sum_{i=1}^{n} wi^2}{(\sum_{i=1}^{n} w_i)^2}$, where w_i = final weight for respondent i (Kish, 1965; Ahrendt et al., 2017),

every statistic and specific set of respondent sub-populations being analysed. Thus analysts should never assume that a sample has sufficient power for the statistical tests they want to do without testing it with the actual data themselves. Rather than report these statistics in this assessment, therefore, we urge data users to use appropriate tools in their specific analyses, considering the country-level output analyses and overall assessment of processes available in this assessment, an approach similar to that found in quality assessments of comparable 3MC surveys as well as other quality frameworks (Beullens et al., 2014a, 2014b; International Monetary Fund, 2012; Eurostat, 2017),

Interpretations of data quality can also be impacted by the extent to which interviewer attitudes and behaviour are considered in analysis (Johnson & Parsons, 1994; Heeb & Gmel, 2001; Benstead, 2014; Benstead and Malouche, 2015; Mneimneh et al. 2017), The 4th EQLS collected interviewer-level sociodemographic data. However, as we discuss in Section 4.3, the lack of a partially-interpenetrated fieldwork assignment process has resulted in the inability to separate interviewer effects from geographic effects in the majority of PSUs across participating countries. We provide recommendations to overcome this challenge in Section 6.3, but because of confounding do not include an analysis of interviewer variance and effects in this assessment.

Survey output quality can also be assessed in terms of both item nonresponse and straight-lining analysis, which indicates possible satisficing in specific batteries of questions. Kantar Public conducted both of these analyses prior to this assessment, examining occurrence by interviewer and identifying relatively minor levels of both item nonresponse and straight-lining.[11] In our review of the report, there were no emergent patterns indicating a challenge to overall data quality.

[11] See the Data editing and cleaning report.

4. Comparative assessment based on current best practices

In the following section, we assess the processes of the 4th EQLS against best practices in the survey research industry, particularly as applied to 3MC contexts. We first define for each of the four phases of the survey lifecycle a set of 3MC survey best practice guidelines, considering both the processes of other major 3MC surveys in the European context, including ESS, Survey of Health, Ageing, and Retirement in Europe (SHARE), and the Programme for the International Assessment of Adult Competencies (PIAAC), as well as survey methodology literature, to support the inclusion of each specific standard in our suggested framework.

We then consider the process followed in relation to each guideline by the ESS, which of the 3MC surveys in the European context is most similar to the EQLS both in terms of questionnaire content and methodology, in contrast to the EU-SILC, which implements output harmonisation, and SHARE, which is a panel survey, although we recognise that the ESS shares neither the organisational structure nor number or breadth of countries with the EQLS. A recent external report of the impact of the ESS notes its synergy with other 3MC surveys in the European context and the extent to which it is considered a benchmark in the industry (Kolarz et al., 2017), While we withhold comment regarding the validity of this claim as it applies to all processes followed by the ESS, we acknowledge the dearth of comparable surveys, the extensive reach, impact, and use of the ESS data, and the relative value for both stakeholders and data users in understanding how the two surveys compare to each other in selecting it as a point of reference to the EQLS.[12] Consideration of ESS practices is followed by that of the EQLS, and discussion of each guideline concludes with a statement about overall compliance with the best practice standard in the 4th EQLS.

4.1 Sampling frame development

To meet EQLS's objectives, the survey sample in each participating country must be representative of the target population, with each element of the population having a known and non-zero chance of being selected. In this type of sample – a probability sample – every person has a chance to be included in the study. Probability samples make it possible to not only compare the sample to the population, but also to compare a sample from one population to a sample from another population, such as one country to another. When developing a sampling frame for 3MC surveys, absolute comparability at all stages of the sampling process is not a requirement for ensuring ultimate cross-national comparability at the data analysis stage (Heeringa & O'Muircheartaigh, 2010), Even in the EQLS, which is organised and funded centrally, circumstances at the local level dictate that flexibility in at least some aspects of sample design is necessary, and the optimal sample design and sample size for one country may not be the optimal design for another. Lynn et al. (2007) suggest that at least two criteria be met in a 3MC survey: (a) study populations in all countries must be equivalent and (b) 'sample-based estimates must have known and appropriate precision in each nation' (p. 108) - meaning, the sample design is transparent, with sample detail available at each selection stage, and that, hopefully, a minimum and comparable precision

[12] When documenting ESS processes, we refer specifically to those implemented in the most recent wave of data collection – ESS Round 8 – unless otherwise noted.

requirement is met in each study country. We build on these criteria when outlining and discussing the following best practices for sampling frame development in a 3MC context.

Comparable target and survey populations must be defined and documented for each participating country.

It is important to develop a detailed, concise definition of the target and survey population, including the elements of the population and the time frame of the group, in order to ensure that each participating country in a 3MC collects data from a comparable population (Groves et al., 2009), Without a precise definition, countries may differentially exclude specific subgroups and differences in the target population may influence estimates of key statistics across countries. The target population may be further refined by definition of the survey population, based on cost, security, or access, although in some cases no further refinement is necessary and the target and final survey populations are the same (Groves et al., 2009), A precise definition of both the target and survey population should be fully documented.

The ESS defines the target population as all persons aged 15 and over who reside in a private household, regardless of nationality, citizenship or language, in each country (ESS Sampling Expert Panel, 2016), Geographic areas excluded from sample populations in specific countries are documented, as are other particular subgroups, such as people who participated in specific recent surveys, as in Finland (European Social Survey, 2017a),

The target population in the 4th EQLS is all persons aged 18 and over whose usual place of residence is in the territory of the participating country included in the survey and who lived in a private household, and further defined by some geographic exclusions.[13] In addition, Kantar Public's documentation noted residency and language eligibility criteria.[14] However, there are no variables in the dataset indicating household members' resident status or language abilities, and the introduction read by the interviewer, per the source questionnaire, did not note this restriction either. Although the initial processes related to defining target and survey populations in the 4th EQLS are comparable to those of the ESS, and comply with best practices, further documentation about how the implementation of processes used per the target population definition are absent.

Sampling frames in each participating country must be identified and evaluated with consideration given to the accuracy of available frames.

Ideally, a sampling frame contains all of the elements of the target population, although in practice few such sampling frames exist. The goal, then, is to choose a sampling frame or a set of sampling frames, either at the individual or the household level, that approximates full coverage of the target population through access to the largest number of elements in the target population, while containing the fewest number of ineligible elements, given the constraints of the survey budget (Groves, 2004), An ideal sampling frame would be fully up-to-date, but real-world limitations often result in at least some level of both undercoverage and overcoverage, duplication, and other inefficiencies. In a 3MC survey, the best approach then is to review available frames in each country, evaluate accuracy, and assess whether one

[13] See Kantar Public's Technical & Fieldwork Report for the 4th EQLS.

[14] See Kantar Public's Sample Evaluation, Enumeration, and Weighting Report for the 4th EQLS.

frame is preferable to another, or if no suitable frame exists and one must be developed (see Best Practice 3 below),

The ESS requires each participating country to evaluate possible sampling frames for accuracy. Where a sampling frame of individuals is not available, or lacks sufficient coverage, countries may use a sampling frame of households or of addresses from which an eligible individual is selected (ESS Sampling Expert Panel, 2016),

Potential sampling frames were reviewed in all participating countries for the 4th EQLS, with respondents selected from individual-level registers when possible. If such registers were available but accuracy was determined to be low, these registers were used to select households, with subsequent selection of respondents at the household level. If individual registers were not available, a household register was used as a sampling frame if available and accuracy was determined to be high. Available documentation indicates that the 4th EQLS followed the process to select sampling frames as outlined by best practices and is comparable to that of the ESS. However, specific register sources were initially missing from available documentation and following best practice should be included in a table in the Sampling and Weighting Report, as well as the date of extraction of data for the sampling frame.

In the absence of an existing sampling frame meeting accuracy criteria, a sampling frame best covering the target population, given budget constraints, must be developed.

Not every participating country in a 3MC survey will have a sampling frame that is both accessible and meets the criteria of accuracy outlined in the previous guideline. Many texts and documents provide detailed guidance regarding the development of area probability samples (Kish, 1965; Üstun et al., 2005), A simple two-stage area probability sample of households, used in many 3MC surveys, involves the following steps: 1) create a list PSUs based on geographic clusters; 2) select a sample of PSUs using a probability sampling method; 3) list all housing units within selected PSUs; and 4) select a random sample of housing units in each listed PSU (Hubbard et al., 2016),

The ESS specifications include instructions regarding those countries where an area sample based on a listing procedure (i.e., enumeration) is applied in the absence of an adequate, pre-existing sampling frame. Specifically, the ESS requires that in these countries, at least twice as many addresses as needed for the gross sample are pre-listed before commencement of fieldwork. In addition, the ESS prohibits interviewers from doing both the listing and the interview itself (European Social Survey, 2015),

An adequate sampling frame was unavailable in 18 countries in the 4th EQLS, with enumeration subsequently utilised.[15] Enumerators in each participating country were given a standardised, detailed instruction manual indicating the process to follow and accompanied by photos and diagrams to further facilitate the enumerator's task. As in the ESS, those responsible for enumeration were not also responsible for interviewing. The 4th EQLS implemented an additional quality control step, where a minimum of 10% of recorded addresses were visually verified, a process which decreases the possibility

[15] Four countries using enumeration in the 3rd EQLS switched to a register in 4th EQLS (Estonia, Spain, Lithuania, and Slovakia), while two countries changed from a register to enumeration (Hungary, Luxembourg),

of error. Available documentation indicates that the 4th EQLS followed the process as outlined by best practices and is comparable to that of the ESS.

If the sampling frame is at the level of a household, then a procedure to randomly select elements from the sampling frame must be determined.

Many sampling frames define population elements at the household level, resulting in a need for a systematic, unbiased method to select a respondent within the household. While various methods are utilised across different surveys, the Kish method is considered to be the ‘gold standard’ for within-household respondent selection in interviewer-administered surveys (Koch, 2018; Lavrakas, 2008; Kish, 1965), With the increased use of technology-driven survey data collection instruments, electronic capture of the household roster can employ Kish-inspired principles to randomly select respondents. Both the original and simulated Kish methods require eligible persons in a household to be listed, with one person subsequently sampled with equal probability from all eligible persons. Other oft-used methods such as the ‘last/most recent birthday’ method may be easier for the interviewer to administer, but have been found to be prone to seasonal biases (Forsman, 1993; Statistisches Bundesamt, 2012), informant’s lack of knowledge (Rizzo et al., 2004), and selection errors (Lavrakas, 2008),

Although the first seven rounds of the ESS permitted use of either the Kish method or the last/next birthday method, Round 8 specifications included explicit instructions to use the Kish method in participating countries when a register of individuals was not used as a sampling frame (Koch, 2018; European Social Survey, 2016a), However, a review of country-specific processes reveals that in several countries, the next/last birthday method (e.g., Czech Republic, Israel, Netherlands) was used to select a respondent from a selected household (European Social Survey, 2017a),

As discussed in Section 4.3, all participating countries in the 4th EQLS use a standardised CAPI instrument, which includes a household roster. With the exception of those few countries where an individual register was used to select respondents, the script in the CAPI instrument was programmed to randomly select a respondent from among eligible members in a simulation of the Kish method. The 4th EQLS fully complied with the best practice of implementing a standardised method of respondent selection across all participating countries, leading to less variation in potential error.

The sample size necessary to meet the desired level of precision must be determined for each participating country.

There are several components to determining the gross and net sample size for each participating country, including the desired level of precision for the statistic(s) of interest, estimates of the statistic of interest from previous surveys, estimated response rates of the survey, and the budget. Design effects also have to be considered in determination of the required net and gross sample sizes, with the target design effect generally ranging from 1 to 3 (Shackman, 2001), Calculation of the effective sample size is subsequently based upon the sampling design characteristics (i.e., extent of clustering, oversampling, etc.),[16] In the European context, there is significant variation in such factors as population size, homogeneity, cost of data collection, and estimated response rates across countries, meaning that the sample size necessary to achieve comparable levels of precision will vary across countries. Specific procedural steps for determining sample size are widely available (e.g., Groves et al., 2009; Hubbard, 2016; Kish, 1965),

The ESS sets a minimum 'effective achieved sample size' of 1,500 in those countries with a population (aged 15+) greater than 2 million, and 800 in those with fewer than 2 million. Each country determined the size of its gross sample considering the realistic predicted impact of clustering, variation in inclusion probabilities (if applicable), eligibility rates (where appropriate), and response rate (European Social Survey, 2015),

The Terms of Reference for the 4th EQLS set the target sample size at between 1,000 and 2,000 in all participating countries.[17] The ultimate gross sample size for each participating country was calculated over the course of several sample replicate releases – termed 'batches' in the EQLS – using estimated response rates and the estimated proportion of eligible households. Available documentation for the 4th EQLS does not, however, indicate the consideration of design effects or calculation of effective sample size in determining target sample size. Recommendations for standardised computation of effective sample size, from a cost/benefit perspective, are provided in Section 6.1.

4.2 Questionnaire development & advance translation, cognitive testing, and translation

Research question(s) or objective(s) must be clearly defined.

Many essential texts and handbooks on survey research identify defining the research questions or objectives as the first step when designing a survey (Fowler, 1995; Groves et al. 2009; Biemer and Lyberg, 2003; Cibelli Hibben et al., 2016), As Biemer and Lyberg (2003) note, a well-specified set of research objectives or questions is a critical component of the survey process. The research objectives define the kind of information that is needed from the survey and are needed to inform many subsequent survey and questionnaire design decisions.

Information about the underlying research objectives is publicly available for questions added for each round of the ESS. For each round of the ESS, the source questionnaire typically consists of one module that is completely new (about 30 items) and in the other rotating module, which is usually repeated and

[16] Effective sample size is defined as the simple random sample (SRS) that would result in the same sampling variance as achieved by the actual sample for any given statistic.

[17] The target sample size was 1,000 in the majority of participating countries, with the following exceptions: Romania (1,100), Poland (1,200), Spain (1,300), Italy (1,400), France (1,500), U.K. (1,600), and Germany and Turkey (2,000),

consists of approximately 10 new items and 20 repeated items. Multi-national teams of researchers are selected to contribute to the design of two rotating modules for the rotating section of the questionnaire. The proposals for the selected rotating modules, which outline the research objectives for the proposed questions, are available online.[18] Detailed documentation is also available online for new items added to the core module including the question aims.[19]

As described in Section 3.1, the 4th EQLS included a number of new questions, most focusing on the quality of public services, such as health-care, long-term care, childcare and other public services. It is not clear whether specific research questions or objectives were defined for the new questions based on information that is available publicly or that we have had access to for the assessment. As in the case of the ESS, research questions or objectives are frequently outlined in proposals. Specific research questions or objectives may not have been defined as such for the new 4th EQLS questions if a proposal was not required as part of the selection process for new questions. We find that the best practice of clearly defining research questions or objectives was not met for the 4th EQLS, based on available documentation.

Subject-area experts, area/cultural specialists, linguistic experts, and survey research experts should be a part of the questionnaire development team or process.

The production of a valid and reliable instrument for a 3MC study requires an array of expertise (Harkness, et al., 2010), Specifically, the skills and abilities for good 3MC questionnaire design are needed from subject-area experts, area/cultural specialists, linguistic experts, and survey research experts (Mohler, 2006), Those involved in the questionnaire development process may not necessarily all be part of one centralised team but should each contribute at various points during the process. For example, specific and short-term input might be needed from experts on a substantive area in the questionnaire or outside contractors may be brought in to provide expertise in particular areas. Collaborators and subject-area experts should be recruited to the extent possible from the different populations represented in the survey to ensure the availability of expertise on given topics as well as critical local knowledge (Harkness et al., 2016),

The questionnaire development process for the ESS incorporates expertise from subject matter experts, area/cultural specialists, translators and translation scholars, as well as survey researchers. The ESS questionnaire is the responsibility of the Core Scientific Team (CST), which is made up of eight institutions. Rotating modules are developed by the Question Module Design Teams (QDTs) together with a sub-group of the CST including experts in cross-cultural questionnaire design selected from different fields. The QDTs include members of the research teams whose proposals were selected for the rotating modules and are able to contribute subject-matter expertise. Since ESS Round 4, the CST has undertaken a consultation process with participating countries on a number of items whose answer categories need to be adapted to the national context before the start of the data collection. A team

[18] See http://www.europeansocialsurvey.org/methodology/ess_methodology/source_questionnaire/

[19] See http://www.europeansocialsurvey.org/methodology/ess_methodology/source_questionnaire/source_questionnaire_development.html

composed of experienced survey translators or linguists and survey researchers carry out advance translation for new or modified questions. Further, members of the CST and the national teams carry out an expert review.

For new questions added to the 4th EQLS, internal documentation indicated that Eurofound engaged subject-area experts to help determine the various dimensions of the concepts that should be measured in the area of public services.[20] Pretesting activities were completed by two different contractors. First, Ipsos MORI carried out focus groups, cognitive interviewing, and an early pilot to help refine and improve the new questions. Further cognitive testing was carried out by Kantar Public in the UK. In-house researchers at Eurofound reviewed and drew on general recommendations from the SQP to refine the questionnaire. Linguistic experts were involved in a translatability assessment conducted by cApStAn, as discussed in more detail below.

Overall, the 4th EQLS questionnaire development process was largely in-line with the best practice of drawing on expertise in most areas. However, it does not appear that a survey methodologist trained specifically in questionnaire design in a 3MC context was involved either throughout the process or at key stages of the questionnaire development process. We also have no evidence for the extent to which consultation occurred with area/cultural specialists or subject-area experts from a broad selection of the different populations represented in the survey. Including area/cultural experts could enhance the questionnaire development process for future rounds of the EQLS.

A translatability assessment or, ideally, an advance translation process should be carried out to make the source questionnaire as easy as possible to translate into other languages and to implement in other cultures.

The wording and content of a source questionnaire plays an essential role in the quality of the resulting questionnaire translations. For this reason, as Smith (2004) has argued, ‘achieving optimal translations begins at the design stage’ (pp. 447), Therefore, it is increasingly considered best practice to undergo a process to check whether the text will be easy to translate. A translatability assessment (TA) and an advance translation (AT) are increasingly applied for this purpose and both have been found to be effective at enhancing the translatability and the cultural adaptability of source questionnaires for translation into multiple target languages (Conway et al., 2014; Dorer, 2015),

While TA and AT share many features, there is one crucial difference. TA relies on the work of individual translators to translate the source questionnaire into several languages. Comments on each individual language from these translators, who typically have had training specifically as translators, are then merged into one common file. AT, on the other hand, involves translations that follow a team approach: interdisciplinary teams composed of translators and survey researchers apply a multi-step translation approach and discuss both the translators’ and the survey researchers’ comments in a review session. In a final step, all comments are not only copied together by one person but discussed and agreed by all actors participating in a review discussion. In this way, AT offers a clear advantage for survey

[20] Internal documentation consisting of meeting minutes from a meeting in Brussels was provided by Eurofound to the University of Michigan.

questionnaire development. However, one drawback of the AT approach tends to be cost because a team of at least three experts needs to be paid for each language assessed.

In the ESS, advance translation has been carried out in three to five languages per round, depending on budgetary constraints. Languages are selected so as to cover as many different language families as possible (Dorer, 2015), The advance translation teams are asked to describe the issues by selecting from among a list of pre-defined categories and by providing comments in their own words. A more detailed description of the advance translation process utilised in the ESS can be found in Dorer (2011) and Dorer (2015),

As noted above, a translatability assessment (TA) was carried out for the 4th EQLS by the service provider cApStAn. For this exercise, translations into the following languages were carried out: Czech, Dutch, French, Polish, Swedish and Italian. This is a good selection of languages because it covers the major language groups of EU countries – Slavic (Czech, Polish), Romance (French, Italian), Germanic (Dutch, Swedish) – as well as the different geographical regions – Eastern Europe (Czech, Polish), Northern Europe (Swedish), Southern Europe (Italian) and Western Europe (Dutch, French), However, it is not clear which language versions were covered for Dutch and French: The Netherlands or Belgium (Dutch) and France, Luxemburg or Belgium (French),

According to our review of the available documentation, while 105 questions were included in the assessment, some languages appear to have covered only part of this number: On the TA worksheets for Czech, French and Swedish, the comments stop at item Q63. Ideally, a TA should include all questions for each language. It is possible that a decision was made that some items were not relevant for some languages, but this decision and the reasoning behind it should be documented.

The excel file we reviewed containing the TA findings and feedback is clear, consistent and easy to understand. The number and level of detail of the comments seem appropriate given the difficulty of the task and the percentage of new text subject to commenting. In general, the comments made in the template are well-formulated, relevant with regard to the questionnaire items and also feasible to address. All comments and subsequent communications were made in English, so the comments as well as the follow-up on the TA comments can be understood even without knowledge of all of the languages included in the assessment. Further, it is useful that all of this information is contained in one file.[21]

However, a detailed review of the comments and recommendations suggests that a survey research perspective could have been helpful. One example is a recommendation in HH3: the answer category ‘in education (at school, university, etc.)’ is recommended to be translated in the sense of ‘schooled’, which may be correct in some languages; however, as a general guideline provided to all languages this seems to narrow down the answer category as compared to the source questionnaire because ‘schooling’ in many countries only determines one fraction of ‘education’. Item Q5, ‘your local area’, offers another example.

[21] Specifically, the different language versions are included in separate worksheets in a single file, and additional information is included on other worksheets: the explanation of the Translatability Categories, the ‘Item by Item Guidelines’ resulting from the TA, as well as information from the Cognitive Interviewing.

The recommendation based on the TA was to translate this phrase in the sense of ‘the place where you live’. However, the word ‘place’ in several languages refers to a relatively small unit, such as a dwelling or apartment or even only a room, and without further instruction, comparability to the larger delimitation of an ‘area’ is not guaranteed.

A lack of documentation made it difficult to assess all aspects of the TA. For example, information on the skills and qualifications of those involved or the instructions or training they received was not available. Based on the documentation available, it was also not possible to determine which comments resulted in changes in the source questionnaire and which ones resulted in translation notes for the translating teams. Missing also is a synthesis of the comments from all languages and information about how and why (or why not) the TA comments were included in the finalisation process of the source questionnaire. There is also no documentation about how the TA process was organised by the contractor or how queries from the translators and/or the contractor towards Eurofound were handled.

Overall, it is our assessment that the TA process for the 4th EQLS was carried out in a well-structured manner resulting in a number of useful comments in six different languages. However, as the decision-making process is not clearly documented it is not easy to assess how exactly the comments were followed-up and why some comments were incorporated and others not.

In addition, as it seems that the TA was carried out after the source questionnaire was finalised, meaning that one of the main ideas behind testing the translatability of a source instrument –modifying the source text accordingly – could not be applied. Carrying out the TA or AT earlier in the questionnaire development and translation process would make it possible to take fuller advantage of the potential benefits of a TA or AT.[22]

An analysis plan should be produced relating each survey question to one or more of the research questions.

Best practice in the development of survey questions also includes producing an analysis plan that outlines how the data will be used (Fowler, 1995; Biemer & Lyberg 2003; Jann & Hinz, 2016), The analysis plan should link each new survey question to at least one research question and detail the data elements that will be produced and how they can be used analytically to address the research question. This process helps avoid superfluous questions that may place undue burden on respondents and ensures that all questions necessary to address the research objectives are included in the questionnaire. A good analysis plan helps focus and guide the question design process.

Information about the underlying research objectives is publicly available for questions added for each round of the ESS, with the source questionnaire consisting of a core section primarily of repeat items and a rotating section comprised of new questions. Multi-national teams of researchers are selected to contribute to the design of two rotating modules for the rotating section of the questionnaire. The

[22] Our assessment of the translatability assessment and the translation process as a whole was informed by a detailed review carried out by assessment team member Brita Dorer, a translation expert.

proposals for the selected rotating modules, which outline the research objectives for the proposed questions, are available online. Detailed documentation is also available online for new items added to the core module including the question aims.

It is not clear whether an analysis plan was developed for the questions added in the 4th EQLS based on available documentation. Similar to research questions or objectives, as discussed above, it is common for proposals for a new survey or new questions to include an analysis plan. As with clearly defined research objectives noted above, an analysis plan may not have been developed as such for the new 4th EQLS questions if a proposal was not required as part of the selection process for new questions.

A team translation approach, ideally following the TRAPD model, should be followed to translate the source questionnaire into target languages.

Current best practice for translation in 3MC surveys is for team translations such as the TRAPD team translation model, the essential elements of which are as follows:

- Translation: At least two independent translations are carried out by qualified translators with at least one of these parallel translations produced by a trained and/or professional translator and all translators experienced in questionnaire translation.
- Review: All questionnaire items should be discussed in a joint review meeting including all translators plus a reviewer and, if possible, an adjudicator. Remote meetings should only be accepted if in-person meetings cannot be organised for practical reasons.
- Adjudication: The final version should be adjudicated, that is signed-off, by an adjudicator.
- Pretesting: A pre-final version of the questionnaire should be pretested with a sample of the final target population. If possible, both quantitative and qualitative pretesting methods should be applied among a sample of the target population in each language.
- Documentation: The entire process should be documented including each step of the process, the people involved and their qualifications and experience, intermediate versions, comments, queries, and potential problems and how they were resolved. Documentation can be done using Word or Excel files or more elaborate tools.

For more detail on TRAPD see Harkness et al. (2010) and Survey Research Center (2016),

The ESS source questionnaire is designed in British English and then translated by each national team. Translations are carried out for any language spoken by at least 5% of the population as first language, and the TRAPD process is carried out for each of these language versions. All national teams are provided with detailed translation guidelines and a translation quality checklist which outlines the procedures to be followed (European Social Survey, 2016c), Verification is carried out by an external service provider to assess the quality of all translations. Translations are also reviewed for linguistic quality and assessed using the Survey Quality Predictor (SQP) to make sure that pre-defined formal criteria are in line with the English source questionnaire. National pretests (P in the TRAPD scheme) are conducted such that any findings can still be incorporated before the national survey instruments are finalised. So far, the quantitative pretests are mainly carried out at national level, but the ESS is moving towards a more qualitative focus for its national pretests. The documentation of the whole ESS translation process takes place in the excel-based '(Translation and) Verification Follow-up Forms, (T)VFFs' at the national level and in Word and PDF files at project level where, e.g., translation queries and answers are listed.

For the 4th EQLS, translations were done from English to national languages for 29 languages. The questionnaire was adapted (in five cases) or harmonised (in seven cases) for languages that are spoken in more than one country. According to the Translation Report, the TRAPD model was used for translating the 4th EQLS source questionnaire. A proofreading exercise was also carried out by a third translator, after which the adjudicator reviewed the translation for a final time in order to produce a final target questionnaire. To assess the translation process for new questions added in the 4th EQLS, we reviewed the Translation Report, the adjudication files for a selection of languages (French (Belgium), French (France), German (Austria), and Italian), and the CVs from all translators and adjudicators.

The CVs reviewed from all translators and adjudicators indicate that the selection of people involved in the translation process was carefully done. All of the translators and adjudicators had either a linguistic/translation or a sociological or survey methodology background and several years of experience in translating and signing off on questionnaires for cross-national survey studies in comparative fields. In this way, the translation process adheres to a ‘team’ or ‘committee approach’ because a team of experts from different backgrounds – mainly linguistic, translation expertise as well as sociological and/or survey methodology – was involved in producing the translations.

Review sessions were carried out remotely while it is ideal for all meetings to be held in-person as this facilitates the intensive discussion and collaboration that is the hallmark of the team translation approach. It is also not clear that a final uniform translation quality assessment or verification step took place. To achieve a homogeneous level of translation quality, it is important that all translations are subjected to the same reviewing steps.

While the translation process for the 4th EQLS followed a team approach, it is not clear from the Translation Report that the ‘Pretest’ component was carried out. This is a crucial element of the TRAPD scheme (the ‘P’ in this acronym) as it is the only step where the pre-final translations can be tested with samples of the target populations in each country. However, it is possible that the pilot tests fulfil the ‘Pretest’ component in the TRAPD scheme because, among other objectives for the pilot tests detailed in the Pilot report, the translated questionnaires were tested, and comments made about the translation quality.[23] Furthermore, several translation problems were detected and subsequently addressed in some languages. This was particularly the case for the Maltese language version.

Nonetheless, while reference is made to the pilot tests in the Translation Report, it is not clear whether the pilot tests were understood and carried out with the same intended goals for the pretests in the TRAPD scheme. For instance, no mention is made to pretesting or pilot testing in the graph outlining the ‘4th EQLS Translation Process’ (pp. 6), We are therefore not able to determine whether the ‘TRAPD’ model was carried out in its entirety.

[23] Eurofound has confirmed that the pilot, carried out in all countries before fieldwork, served the purpose of a final pretesting.

Other aspects of the translation process for the 4th EQLS are well documented. The translation templates were well-structured and organised for capturing the entire decision-making process from the initial parallel translations until the final agreed version. We note only that the documentation could be more comprehensive, and several elements made more clearly identifiable. For example, the terminology used is not always clear and, as it does not readily correspond to the classic TRAPD approach, should be better explained; examples are the terms 'adaptation', 'harmonisation', 'third adjudicator', etc.

In sum, we believe that the translation process was largely in line with current best practice and that the translated questionnaires for the 4th EQLS are of sufficient quality to produce comparable data between all countries participating in the survey. Recommended changes, detailed in the recommendations section, would further enhance the quality and outcomes of the translation process.

An appropriate set of pretesting and/or post-hoc evaluation methods should be used to assess the quality of questions, based on available resources.

Pretesting plays an essential role in identifying and potentially reducing measurement error that damages statistical estimates at the population level and can threaten comparability across populations in a 3MC study (Caspar et al., 2017), While most researchers would agree that some degree of pretesting should be done in each study country before a 3MC survey is fielded, no standard currently exists for which type, combination, or amount of pretesting that should be done.

Pretesting techniques typically used in single population surveys, such as cognitive interviews, focus groups, expert reviews, pilot studies, and behaviour coding, among others (Presser et al., 2004), can also be effectively applied in 3MC surveys (Caspar et al., 2017), Different types of pretesting techniques tend to yield different types of results and no one technique can offer a comprehensive set of findings about the quality of or potential problems with a questionnaire. It is therefore ideal to combine pretesting techniques and/or post-hoc evaluation methods in a way that takes advantage of the strengths and minimises the weaknesses of each method.

While there are a number of challenges in designing and implementing pretests for cross-national studies and time and resource limitations are typical constraints (see Pennell et al., 2017 for discussion), minimum best practice is to carry at some form of pretesting in each study country. This may take the form of a pilot test during which field procedures are also tested. Further, an expert review of the questionnaire is relatively low cost and should be considered standard practice as well as cognitive interviewing in the source language at a minimum, with additional language families, major regional subgroups, or countries added as resources permit. Results from a recent study examining how question evaluation methods compare in predicting problems suggests that the best combination of methods may be expert reviews followed by cognitive interviews (Maitland and Presser, 2017), lending support for this best practice. In addition to identifying problems, in the 3MC context, cognitive interviewing should be carried out not only to identify potential problems with questions but to investigate the constructs captured by questions (i.e., construct validity) and whether or not those same constructs are captured across various groups of respondents (Miller, 2018), In this way, as Miller (2018) notes, cognitive interviewing can help ensure validity and comparability, which is particularly essential for studies seeking to make comparative estimates (see also Braun, et al., 2015), Maitland and Presser also found computer-based methods (SQP and the Question Understanding Aid (QUAID)) to be the least predictive of question problems, which suggests that these methods should be used to complement but not to replace expert reviews and cognitive interviewing. Additional pretesting approaches and post-hoc evaluation should also be carried out to the extent possible based on available resources.

The questionnaire development process for the ESS incorporates a number of pretesting methods as well as a post-hoc evaluation carried out for each round using Multitrait-Multimethod (MTMM), In the ESS, there are two types of pretests: (a) during the questionnaire design phase, organised by the CST: methods that have been used are expert review from members of the CST), the SQP, cognitive interviewing, advance translation and quantitative testing on omnibus surveys and in a two-nation pilot survey[24]; and (b) during the translation phase organised by the national teams ('P' in the TRAPD scheme), It is mandatory that each participating country carry out a pretest to, at a minimum, check whether the translations of the questionnaire are consistent with the intended meaning, and whether CAPI/PAPI routings work properly. However, countries vary widely in how the pretests are carried out. For example, ESS Round 7 saw substantial variation in pretesting in terms of timing (how long before fieldwork was set to begin), the number of completed interviews, the mode of interviews, whether cognitive interviewing was done, and whether or not interviews were recorded (Beullens et al. 2014b), As Beullens et al. (2014b) report, only two countries used cognitive interviewing (Estonia and Switzerland), four countries skipped checking the translation from English into the national language, and recording was used only in France (audio and video), Lithuania and Portugal, and in Belgium (video only), Since Round 1, the ESS has also incorporated MTMM experiments to estimate the measurement quality of questions (see Saris and Gallhofer [2014] for discussion),

Pretesting for the 4th EQLS largely meets minimum and in some ways exceeds best practice for 3MC surveys. Meeting best practice, pilot tests were conducted in all study countries to test all aspects of the questionnaire as well as survey administration, contact procedures and interviewer instructions and cognitive testing was carried out on the source questionnaire by Kantar in the UK. Exceeding best practice, early pretesting done by Ipsos for developing the new source questions included focus groups in the United Kingdom and Poland, cognitive interviewing in the United Kingdom, and an early pilot in Belgium and Poland to help refine and improve the new questions. However, no expert review was conducted which may have enhanced the results of the pretesting activities and minimise the types of suggested revisions we noted in Section 3.1. Further, the cognitive interviewing studies that were done could be improved by additional focus on validating question constructs and comparability as well as identifying problems.

Develop a documentation scheme for questionnaire design decisions and changes to the source questionnaire across time for repeat surveys.

Best practice for questionnaire development includes thorough documentation of the process, decisions made, and changes made to the source questionnaire for repeat surveys. The development of indicators and questions should be documented from start to finish (e.g., any modifications made to questions at different stages and why) (Harkness et al., 2016), As Harkness et al. (2016) also note, version control

[24] See http://www.europeansocialsurvey.org/methodology/ess_methodology/source_questionnaire/source_questionnaire_development.html

procedures are essential whenever a source questionnaire is modified across time. Flexible documentation templates should be used to facilitate consistent and well-organised documentation.

For each round of the ESS, changes made to the core module are documented along with the reason for the change with references to more detailed information available in other documents. Successful proposals for the rotating modules document the development of indicators and questions.[25] Question design templates are also available for the rotating modules. Minor changes to the core questionnaire may also be proposed to improve the questionnaire in the different participating countries. These types of changes are often made to categorical variables that are very country specific, such as religion, education status, and income deciles, and are documented by country in the documentation reports. Changes in existing translations are documented in specific reports per round, although this report has so far only been finalised for Round 5. Changes in the actual source questionnaires for each round are available from the ESS ERIC HQ upon request.

New and modified questions for the 4th EQLS are identified in the source questionnaire that is available online.[26] The final source questionnaire also denotes which questions were asked in which years of the survey (e.g., 2003, 2007, 2011, 2016), However, we are not aware of detailed documentation related to the development of new questions or to modifications made to questions at different stages and why.

Show cards should be produced for those survey items as needed, for use by interviewers in all participating countries following a standard protocol.

Show cards are a device used in face-to-face surveys to visually show response options to respondents. In the 3MC context, show cards should contain the exact text used in the response categories in the questionnaire, and should be identical in layout between the translated version and the source version. Without instituting a standardised process, differences can occur between countries and contribute to measurement error in a 3MC survey. A review of previous rounds of the ESS found numerous deviations, including: (a) differential inclusion of the start of the response sentence on show cards; (b) show cards putting the answer codes in boxes, omitting the numbering of the categories, or drawing arrows to indicate the end points, unlike in the original; and (c) survey items that were formatted as single questions, each with their own answer scale, rather than formatted as batteries of items (Dorer, 2012), Ideally, show cards should be a set of physical documents, separate from any technical instrument (i.e., a CAPI instrument), Visually displaying show cards on the CAPI instrument may be understood by the respondent as an invitation to read along with the interviewer, which can introduce measurement error.

In the ESS, interviewers are provided with show cards, and the central coordinating centre recommends that show cards are laminated and placed in a binder to facilitate the interviewer's work. Instructions also note that the show cards should only contain the response options, and neither the text of the question nor options for 'don't know' and 'refuse', with exceptions to this clearly indicated in the questionnaire

[25] See http://www.europeansocialsurvey.org/methodology/ess_methodology/source_questionnaire

[26] See https://www.eurofound.europa.eu/sites/default/files/ef_survey/field_ef_documents/4th_eqls_final_master_source_questionnaire_12_june_2017_-_updated_07_september_2017.pdf

(European Social Survey, 2016b), Interviewer instructions note that interviewers should choose a seating arrangement so that respondents cannot see the computer screen (or paper questionnaire) (ESS – Interviewer Manual),

In the 4th EQLS, separate show cards were not used. Interviewers were instructed to show response categories to the respondents on the tablet screens for particular questions, as indicated in the Source Questionnaire. However, the Source Questionnaire itself contains inconsistent wording to refer to show cards, and implies in some questions (e.g., HH2) that there is a physical show card available. Further discussion with Kantar Public provided additional information about the physical layout on the screen and how the response options 'don't know' and 'refuse' appeared, but screen shots for each survey question or battery of items is not available to data users. The process followed in the 4th EQLS does not comply with the best practice of use of standardised, physical show cards, which provide analysts with a more comprehensive understanding of the context of the question/answer process.

4.3 Fieldwork (Implementation, monitoring, contact procedures, nonresponse, and paradata)

Fieldwork encompasses a number of activities related to data collection: development of comprehensive interviewer training tools, interviewer recruitment, training, and monitoring, and interview verification. The complex task of collecting comparable data in the context of a 3MC survey necessitates researchers to follow a number of best practices related to data collection.

A standard CAPI instrument, which includes components for both data collection and sample management should be used in all participating countries.

As technology becomes more accessible and affordable and its use increases worldwide, computer-assisted personal interviewing (CAPI) is increasingly replacing paper-and-pencil instruments (PAPI), The first of two essential components of the CAPI instrument – the sample management system – is used to release sample lines to interviewers and to record interviewer call records and contact attempts. Ideally, it is integrated with the second essential component of the CAPI instrument – the data collection system – which is the mechanism used to administer the actual survey questionnaire. Ideally, the latter includes collection of an audit trail which captures all movement through the questionnaire with associated timestamps. The integrated system greatly enhances the monitoring of interviewer workload and performance. Laptop computers have generally been the instrument of choice for CAPI, but cheaper options such as tablets, smartphones, and other handheld device are increasingly being employed.

There are numerous advantages of electronic capture of data, particularly with regards to increased capacity for paradata collection and quality control (Kreuter, 2013a) and leverage of these is further enhanced in the 3MC context if a central data collection and sample management system is implemented. In addition to this advantage, use of a central, standardised CAPI instrument can reduce measurement error, including the error which could result from cognitive processing, context effects, and interviewer effects due to differences in instrument design and administration. Because of the significant advantage, use of a standard CAPI instrument is the recommended mode of data collection in interviewer-administered, face-to-face 3MC surveys.

Although there was a strong recommendation directed at participating countries in the ESS to use a CAPI instrument, it was not obligatory (European Social Survey, 2015, 2016a), Several countries used PAPI for both sample management and questionnaire administration, while others used a CAPI instrument only for the questionnaire component, and PAPI for the sample management component. Among those who used CAPI, there was no central standardisation employed (European Social Survey, 2017a), SHARE, however, does requires use of the same integrated systems across its participating countries and has done so since its first wave of data collection.

Like SHARE, the 4th EQLS not only required use of a CAPI instrument in all participating countries, the technology used was developed centrally and fully standardised across all countries, with two very minor exceptions (see Section 2.3), The CAPI instrument included both a sample management system and the questionnaire itself. Based on available documentation, the 4th EQLS is fully compliant with best practices, and the use of a standardised system is a critical advance in minimizing error.

A standard CAPI instrument should be developed centrally and then thoroughly evaluated in all participating countries.

The technical design of a survey instrument focuses primarily on the design of the actual software that collects data related to sample management and delivers the questionnaire content, including the format, layout, and other visual aspects of the presentation or context of survey questions (Hansen et al., 2016), In a 3MC survey, programming of the instrument should be developed centrally to maintain standardisation, even though translation(s) in participating countries will by necessity differ. In general, the layout in the source questionnaire should be preserved in subsequent translated versions, resulting in a translated version identical to the original except for the words. As the objective in a 3MC survey is to collect comparable survey data, it is crucial that design implementation does not contribute to measurement error. Testing both at the level of the central coordinating centre, as well in each participating country, is an integral part of the quality assurance process.

As noted above, ESS does not require use of CAPI, and there is currently no effort to develop a standardised instrument, and therefore there is no documentation of relevant standards. Documentation currently available on evaluation includes details at the country-level for whether testing the CAPI script was a component of the pretest, and the extent to which script testing occurred (European Social Survey, 2017a),

While the 4th EQLS developed and deployed a standardised CAPI instrument for data collection, documentation on the development process and subsequent testing is sparse and may be related to issues of competition among data collection firms and reluctance to share possible industry advantages. Recommendations for documentation in future surveys are included in Section 6.3.

A checklist of minimum interviewer candidate requirements should be established and a comprehensive, standardised interviewer training must be conducted.

Interviewers implement the survey design and as such are integral to the data collection process as members of the research team. They are often required to perform multiple tasks with a high level of accuracy. There are a number of important criteria to consider when recruiting interviewers, including previous experience, education, computer skills, language proficiency where applicable, and navigation skills (Stoop et al., 2016; Jäckle et al., 2013; Lipps, 2007; Pickery & Loosveldt, 2000, 2002; Vassallo et al., 2015), In the 3MC context, these criteria should be determined at the level of the central coordinating centre and implemented in a standardised fashion in all participating countries.

In the field, interviewer behaviours can impact sampling error, nonresponse error, and processing error. Interviewers can also contribute to measurement error by influencing responses through their personal attributes and their behaviours, through what is often referred to as ‘interviewer effects’. Surveys in a 3MC context present a particular challenge due to differences in the cultural environment, existing infrastructure, and resources available (Smith, 2007), However, a comprehensive interviewer training, standardised across all participating countries in a 3MC survey, and which includes general interviewer training techniques, study specific training, and certification, is crucial to reducing these effects.

The ESS prefers countries to work with experienced interviewers, and the ESS manual requires interviewers to have experience doing face-to-face interviews and receive both general training on face-to-face interviewing and in-person, study-specific training (European Social Survey, 2016b), Documentation for participating countries indicate that interviewer trainings are be completed in-person, and the ESS coordinating centre developed and distributed an interviewer manual, briefing slides, and a briefing example interview available for use during interviewer training (European Social Survey, 2016b), However, participating countries were not obligated to use these materials, and indeed, there was variation in implementation (European Social Survey, 2017a), This absence of a completely standardised interviewer training increases the possibility of comparison error in the resultant data collection, as interviewers may implement various data collection activities differentially.

The 4th EQLS requirements for the selection of field force were that interviewers must be native speakers of the language used in the country (or part of the country), with at least one year of experience in survey research and that interviewers must have participated in at least three face-to-face social surveys in the past five years. An interviewer manual was developed by Kantar Public in consultation with Eurofound, and was used in all interviewer trainings, protocols for which were standardised across participating countries. Available documentation in the 4th EQLS indicates compliance with best practices for interviewer recruitment and training, thereby maximizing standardisation and decreasing potential comparison error.

The use of incentives for participation should be determined and documented in each participating country.

Nonresponse can be reduced by offering respondents an incentive for participating in a survey (Singer, 2002), In the 3MC context, incentives are likely to vary across participating countries based on local resources, customs, and ethical regulations (Kessler, et al., 2008), and impact may vary as well (van den Brakel et al., 2006), If an incentive is used, the amount and type, time of implementation, and any special strategy, such as increasing the amount of the incentive in the final weeks of the study, should be thoroughly documented.

While nearly all participating countries in the ESS used an incentive, the specific incentive (monetary vs. non-monetary) and timing (conditional vs. unconditional) varied. All use of incentives is fully documented (European Social Survey, 2017a),

In the 4th EQLS, about half of participating countries issued some form of incentive, typically non-monetary. Among those countries providing incentives, the majority were conditional; that is, given to respondents upon completion of the survey. As use of incentives is also fully documented, we find the 4th EQLS to be fully compliant.

A standard pretest protocol should be developed and implemented in each participating country.

Pretesting plays an essential role in identifying and potentially reducing measurement error that affects statistical estimates at the population level and thus endangers comparability in 3MC surveys. Pretesting includes activities designed to evaluate a survey instrument's capacity to collect the desired data and the overall adequacy of the field procedures. Pretesting in all participating countries is crucial, and standard minimum criteria for the pretest should be defined and implemented in all countries, including number of attempted/completed interviews and any quotas of specific subpopulations (e.g., five females and five males within each of three different age categories), number of interviewers to participate, and the timeframe in which to complete the pretest. As noted in Section 4.2, some form of pretesting is recommended as best practice with regards to questionnaire development. While this pretest may be separate from or integrated with the form of pretest recommended here, it is imperative that the criteria noted here be applied.

The initial specifications for Round 8 of the ESS noted, as new for Round 8, a set of pretesting guidelines (European Social Survey, 2015), However, these are not publically available for review. Documentation of country-specific processes followed in Round 8 reveals substantial variation in the number of pretest interviews, with the U.K. completing five pretest interviews, while in Finland, 144 pretest interviews were completed. All pretests were conducted face-to-face, and in some countries pretest interviews were recorded for later assessment (European Social Survey, 2017a),

In the 4th EQLS, pretesting was standardised and took place in the same four weeks across all participating countries. Each country completed about 30 interviews across both urban and rural PSUs and in bilingual countries, 40 interviews were completed, with additional pretests completed in several countries to further investigate low response rates and/or unusual interview length. The standardisation in pretesting increases the potential for comparable reductions in error across countries, and the process implemented in the 4th EQLS is fully compliant with best practices.

Mode of first contact should be standardised across all countries.

An individual's initial contact with an interviewer can determine whether s/he becomes a respondent or a non-respondent. Whether this initial contact takes place in person or by telephone, however, can impact the outcome of the interaction, and investigations of mode effects demonstrate that initial face-to-face contact results in greater contact, and subsequent cooperation, than initial telephone contact (Hox & De Leeuw, 1994; Holbrook et al., 2003), The initial mode of contact should be standardised in 3MC surveys so as not to differentially impact nonresponse.

The ESS allows for first visits to be via telephone or face-to-face contact, with a number of countries selecting telephone as the initial mode of contact (European Social Survey, 2017a),

With the exception of Sweden, all initial contact attempts in the 4th EQLS were initially face-to-face. In Sweden, respondents in the 4th EQLS were selected from an individual register, and telephone numbers were available for the majority of respondents and used for the initial mode of contact, providing a number was available. As noted in Section 3.2, response rates in Sweden were very low and possibly impacted by this differential use of mode. Recommendations in Section 6.3 consider changes to initial contact procedures in future Eurofound-led surveys.

At minimum, a partially-interpenetrated interviewer field assignment plan should be developed and implemented to permit estimation of interviewer effects.

As noted in Best Practice 3 (above), interviewer behaviours and attitudes can impact many sources of error, and this impact can occur differentially across countries (Groves et al., 2009; Blom et al., 2011; Loosveldt & Beullens, 2014; Japec, 2005; Beullens & Loosveldt, 2014, 2016; de Jong et al., 2016; Mneimneh et al., 2017), The cluster design of most area probability sample surveys confounds the sampling and non-sampling (i.e. interviewer) variances if only one interviewer is assigned to a specific cluster. Although such confounding is eliminated if respondents are randomly assigned to interviewers, in practice this is nearly always cost prohibitive. More feasible is the use of interviewing teams with at least two interviewers assigned to each primary sampling unit (PSU), which permits the estimation of measurement error introduced by the interviewer. Known as a “partially-interpenetrated design”, this approach facilitates multi-level modelling in data analysis to estimate interviewer and design effects simultaneously (O’Muircheartaigh & Campanelli, 1998),

The most recent interviewing materials distributed to countries participating in the ESS acknowledge that there is interviewer variance in both process and output data, increasing the likelihood of interviewer-related error (European Social Survey, 2016b), To permit estimation of errors and reduce interviewer variance, the ESS training manual notes that interviewers are permitted to work on a maximum of 48 cases. The ESS coordinating centre also strongly recommends that partial interpenetration is used in field assignment, to decrease confounding between interviewers and geographic areas and allow for a more thorough assessment of interviewer effects (European Social Survey, 2016b; Stoop et al., 2016),

Materials used in the 4th EQLS did provide instructions to participating countries on interviewer assignment regarding interviewer workload (requiring to limit the maximum number of interviews carried out by an interviewer to 20), However, available documentation does not discuss the determination of interviewer assignment to specific PSUs, nor the relationship between interpenetration and the ability to assess interviewer effects. Preliminary analysis of the data demonstrates a great deal of confounding of geographic region and interviewer assignment, meaning that, as in the ESS, interviewer effects cannot be fully assessed. In Section 6.4, we provide recommendations regarding how the EQLS can address this in the future.

Identify both computer-generated and interviewer-generated paradata to be collected and develop clear analysis procedures for the different types of paradata.

The use of paradata to investigate and reduce survey error provides a wide-range of information about the survey data collection process. Nonresponse error, measurement error, coverage error, and sampling error can all threaten the final survey estimates, and these effects are magnified due to comparison error in a 3MC survey (Smith 2011, 2018), With the increasing use of complex technology in surveys, paradata have become widely available to researchers, providing additional tools to evaluate and reduce survey error sources across participating countries (Kreuter, 2013a), Such data can be used from a sample management perspective, such as a mechanism to drive responsive design, where researchers continually monitor selected paradata to inform the error-cost tradeoff in real-time, as the basis for altering design features during the course of data collection or for subsequent waves. Paradata can also be used to monitor and evaluate interviewers during data collection (Kirgis & Lepkowski, 2013; Mneimneh et al., 2018; Hyder et al., 2017) as well as to study interviewer effects and the interview context in the analysis phase of a project (Johnson & Parsons, 1994; Heeb & Gmel, 2001; Benstead, 2014; Benstead & Malouche; 2015; Mneimneh et al. 2017),

The ESS, in the planning and implementation of each round, has built in a feedback loop that includes continuous improvement (Pennell et al., 2017), It closely monitors the survey process, collects various types of paradata and other auxiliary data using contact forms (Stoop et al., 2010), and documents the paradata for each round of the survey. For example, interviewers are expected to collect data to facilitate evaluation of nonresponse bias, including thorough contact data and information about the respondents' home and neighbourhood for all sample cases, regardless of outcome. Patterns in the information about noncontacts, ineligibles, refusals etc. can point to procedures that can help improve the response rates in future ESS rounds or even during fieldwork.

Use of a standardised CAPI instrument in all participating countries in the 4th EQLS facilitated a standardised collection of paradata, as summarised in Appendix 5 (Table A13), Data were collected both about households where an interview was completed, as well as households which resulted in a final outcome code indicating some form of nonresponse. Some details regarding interviewer characteristics were also captured. These paradata could facilitate basic response bias analysis, as we recommend in Section 6.3. GPS data were collected at the enumeration stage for those countries were registries where not available, but documentation is unclear as to whether these data were again collected during the interview, or during a contact attempt, and it appears these data were not used in the field period for monitoring purposes. Documentation indicates that the other paradata were used during the field period to monitor response rates, and after the fieldwork to assess some aspects of interviewer performance. However, as also discussed in Section 6.3, there are additional paradata that the EQLS may consider collecting and utilizing in the future both during fieldwork through responsive design, and afterward in analysis to correct for different sources of error.

A data-driven assessment protocol for the selection and verification of cases should be established and include thorough documentation for both selection rationale and verification outcome.

Non-standardised interviewer behaviour such as reading questions too quickly, skipping questions, not reading the questions as worded, or using improper probes, etc. may introduce measurement error. The increasing use of paradata, such as item level time stamps and behaviour-coding data, enhances the ability to monitor interviewer behaviour and evaluate the data they collect (Durand, 2005; Laflamme & St-Jean, 2011; West & Groves, 2013), Retraining or other interventions can be applied to any interviewers not following protocol. Standardised collection and utilisation of paradata in all participating countries in a 3MC survey could result in reductions in measurement error during data collection.

In the ESS, cases are monitored for correct application of final outcome codes and potential nonresponse bias. The ESS8 Specifications state that outcome codes for 10% of interviews, 5% of refusals, and 5% of noncontacts and ineligibles should be verified, with verifications conducted by an impartial party if possible (European Social Survey, 2016b), ESS documentation provides detailed information about progress monitoring activities such as those relating to outcome codes and interviewer performance, and includes examples of measures to be monitored, how to use the resultant data, and possible actions to be implemented (Koch et al., 2016), Documentation on data verification selection and on the result of the verification process is not available, and real-time intervention remains a challenge for the ESS.

Participating countries in the EQLS were required to back-check at least 10% of completed interviews. Interview(ers) were selected both at random, and by a targeted process of selection of interview(er)s prioritised as high risk, as defined by set criteria.[27] Each interviewer had at least one interview selected for back-checking. The majority of back-checks were conducted by telephone, with a face-to-face if no telephone number was available (and by post only in Germany), The process used in the EQLS complies with best practice in that interview(er)s were targeted in part by pre-determined criteria. However, there is no data available to indicate the specific criteria for which individual interview(er)s were selected, nor is there data related to the outcome of the verification beyond whether the case passed or failed the verification. There are multiple reasons for verification failure (e.g., interviewer error, falsification, etc.) and such data can be informative to both the field process and subsequent quality assessment. Additionally, there is no evidence of verification for those cases assigned refusal, noncontact, or ineligible final outcome codes. Section 6.3 includes further recommendations on how paradata might be used further to identify interviewers at risk of having data quality issues.

A nonresponse bias analysis should be conducted for all participating countries.

Response rates have been declining throughout Europe and elsewhere, with researchers increasingly concerned about the potential impact of nonresponse bias (Brick & Williams, 2013; Groves, 2011; Kreuter, 2013b; Peytchev, 2013), Nonresponse bias includes two components: the response rate and the differences between respondents and non-respondents. If respondents and non-respondents are identical, then there is no nonresponse bias, regardless of the size of the response rate, while if non-respondents differ from respondents, then the response rate is inversely proportional to the resulting bias (Groves, 2006), Available paradata such as call history data and interviewer observations for both respondents and non-respondents may inform researchers about the likely differences between respondents and non-respondents, and help researchers to evaluate nonresponse bias (Kreuter & Olson, 2013; Wagner & Stoop, 2018), In a 3MC survey, where differential nonresponse bias can impact cross-national data comparability, the risk of bias both between and within countries must be assessed (Wagner & Stoop, 2018),

[27] Criteria focused on issues such as contact timing and interview length. See the QAR for additional details.

Aspiring to mitigate one component of nonresponse bias, the ESS sets as targets minimum contact rates and responses rates (97% and 70%, respectively), with an alternate target minimum response rate higher than in the previous round for those countries where a 70% response rate is unlikely, although there is no evidence that target achievement is associated with prioritisation of nonresponse analysis (European Social Survey, 2015), However, ESS does use various paradata in extensive analyses of nonresponse bias (Blom et al., 2011; Billiet et al., 2007; Stoop et al., 2010a, 2010b),

As noted in Best Practice 8, above, the collection of extensive paradata facilitates nonresponse bias analysis. In the initial tender for the 4th EQLS, tenderers were invited to propose possible paradata for this purpose. However, in the available documentation, there is no evidence that a proposal specifically to investigate nonresponse bias was developed, nor is there evidence that this type of analysis has been completed beyond that which was done with regard to development of weights to account for differential response. Section 6.3 provides recommendations on both basic and extensive nonresponse analyses, as well as a suggested threshold at which such analyses might occur.

4.4 Weighting

The construction of survey weights for use in statistical computations is an essential step after data are collected. Weighting adjustments are commonly applied in surveys in a series of stages to compensate for different selection probabilities, under coverage, nonresponse, and to make weighted sample estimates conform to external values (Kalton & Flores-Cervantes, 2003), Survey weights are necessary for producing accurate and comparable population estimates. We identify the following best practices for weighting for 3MC surveys.

The following survey weights should be constructed, as needed, and fully documented:

- Design or base weights to correct for different probabilities of selection.

Survey weights are typically needed to correct for unequal probabilities of selection for complex sample designs. For each respondent, the design weight should be calculated by multiplying the probabilities of selection at each stage of selection (e.g., primary sampling unit (PSU), secondary sampling unit (SSU), the selection of addresses or individuals when using a register, the selection of one respondent in each household (based on the sample design used in each country),

For the ESS, a design weight variable is provided and documentation includes information about the construction of the design weights and application for analyses (European Social Survey 2014, 2017a),

For the 4th EQLS, a final design weight is included in the dataset. The Sampling, Enumeration, and Weighting Report outlines how the design weights were calculated including a summary of the design weight calculation at each stage of selection for each type of sample design.

- Weights to adjust for under coverage, nonresponse, and to make weighted sample estimates conform to external values.

Nonresponse adjustments should compensate for differential response rates for mutually exclusive and exhaustive subgroups in the sample that are related to the statistic of interest. This usually includes different geographical units (e.g., differences in response between regions or between different types of neighbourhoods) and different demographic and socioeconomic characteristics, such as age and sex. Subsequent adjustments should be employed to make sample totals for key demographic and socio-demographic subgroups match population totals available from some external data source separate from the survey, such as population distributions or totals obtained from the sampling frame, a large high-quality survey such as a labour force survey, or a census or other official statistics.

Nonresponse and population weighting are often done in two stages, but the same methods can be applied for each of these purposes (Kalton & Flores-Cervantes, 2003), Methods include cell weighting or raking or model-based approaches such as generalised regression estimation (GREG) or logistic regression weighting.[28] Model-based approaches are increasingly employed as they can incorporate more auxiliary information than is possible with adjustment cell methods. A further advantage of model-based methods is that, when done correctly, the amount that the individual design weights must be adjusted for the final weight can be potentially minimised.

Final post-stratification weights are obtained by adjusting the design weights with the nonresponse and population weights such that they will replicate distributions of the external data source, sometimes called control data. It is important to use the highest quality and most up-to-date control data available for each country. Missing values for any variable needed for post-stratification adjustments should be imputed. Documentation should specify the source of the control data for each country, the method used for post-stratification, the extent of missing data for variables used in weighting, and how missing data was handled. In a 3MC context, it makes sense to apply the same or similar approach to weighting in each country to the extent possible. However, differences in the type and quality of external data sources can present a challenge. For comparability, the key is that weighting adjustments are made based on the best available data in each country. Post-stratification weights for the ESS are constructed using information on variables for age, gender, education, and region primarily from the European Union Labour Force Survey (LFS) and other sources when necessary. Available documentation includes information about the source of the control data and the methodology used for post-stratification (European Social Survey, 2014), Most countries follow a similar approach to missing data with some exceptions, which are clearly identified. Tables with the actual control data for each country, the source available for each country for each survey round, along with the date of extraction are provided for many countries.[29] This information is intended to be updated periodically when new data become available and new post-stratification weights are subsequently released.

The EQLS Sampling, Enumeration, and Weighting Report outlines the approach taken for constructing the post-stratification weights for the 4th EQLS. A table in this report lists the socio-demographic variables and categories used and the sources for their population targets, which included Eurostat, EU-SILC, and local country level data. However, tables with the control data for each country, the source

[28] Valliant et al. (2013) and Kalton & Flores-Cervantes (2003) provide detailed discussion and examples.

[29] See ESS8 Annex A5 Population Statistics ed. 1.0, available for download: http://www.europeansocialsurvey.org/data/download.html?r=8 .

(e.g., URL, database, etc.), and date extracted is not provided. [30] It is also not clear how missing data were handled.

- Supranational or population size weights to adjust for different national population sizes.

In 3MC surveys, supranational or population size weights are needed for analyses that combine data from more than one country. The population size weight makes an adjustment to ensure that each country is represented in proportion to its population size.

The ESS provides population size weights, which can be applied in combination with the design or post-stratification weights. Information about the construction of PWEIGHT and its application for analyses is publically available (European Social Survey, 2014),

For the 4th EQLS, a final calibrated cross-national weight is provided in the final dataset that can be applied when calculating statistical estimates (percentages, averages) and their confidence intervals based on data from two or more countries. Information about how this weight was calculated is provided in the Sample Evaluation, Enumeration, and Weighting Report.

Weight trimming or other methods for addressing widely varying survey weights should be considered, applied as appropriate, and documented.

Extreme weighting adjustments can increase the variances of estimates and cause problems with subgroup analyses. Survey weights should therefore be examined, and methods considered to address widely varying weights. This can be done at earlier stages of the weighting process but should be done at the final stage also. Weight trimming involves establishing an upper cut-off point for large weights, reducing weights larger to the cut-off value (also referred to capping or truncating), and then spreading the weight above the cut-off to the non-trimmed cases (Henry & Valliant, 2012), As Valliant et al. (2013) note, methods for setting the bounds are generally arbitrary and a matter of preference or historical precedence. Other techniques exist for handling or constraining extreme weights, however, weight trimming or other so-called ad-hoc methods are common practice. Whatever approach is taken, the key is that the variation in weights is addressed and that actions taken are applied consistently and are well-documented.

Documentation on weighting for the ESS states that both design and post-stratification weights are scaled to the sample size and truncated around the value of 4 which suggests that an ad-hoc trimming approach is used (European Social Survey, 2014), No further information is available about how the variation in weights is addressed.

Weight trimming was also employed for the 4th EQLS. Trimming procedures, or capping, based on the terminology used in the Sampling, Enumeration, and Weighting Report, was carried out on the design weights based on the individual stages of selection. This was done to correct for some oddities in the enumeration process, for countries where enumeration occurred as part of the sampling process, resulting

[30] Eurofound has subsequently updated the Sample evaluation, enumeration and weighting report.

in some extreme weights. Ideally, these oddities would not have occurred and should be avoided in the future, as we discuss in Section 6.4. However, given that they did occur, it is reasonable that the resultant weights were handled in the way they were. Other design weights at the individual stage were capped to address some extreme weights that were presumably due to large household sizes. Weights at the individual stages could have been capped at the final design weight or the final calibration weight stage but it is not unreasonable that they were handled at these earlier stages. Finally, according to the documentation, all design weights were capped to three times the median value. However, at the final weighting stage, twenty possible final calibration weights for each country were evaluated some of which used the capped design weight and others that used uncapped design weights. Final calibration weights were then selected based on an appropriate set of criteria. Overall, we find that the approach taken to address the variation in weights for the 4th EQLS was reasonable and adequately documented.

A guide should be provided to assist end users with the correct use of survey weights.

Data from complex samples must have the survey weights applied correctly to adjust for such a design in any statistical computations. As such, it is important to provide data users with information about when different weights should be used and how weighting may affect the data. The ESS has produced a specific guide and encourages data users to consult it before conducting any analysis of ESS data (European Social Survey, 2014),

As noted above, the Sample Evaluation, Enumeration, and Weighting Report provides detailed information about the weighting strategy, the construction of survey weights for the 4th EQLS, as well as when the weights should be applied.

If comparison over time is an important goal, the weighting strategy should be kept as consistent as possible for multi-wave surveys or modified weights should be produced and made available based on the most recent weighting methodology.

For the ESS, it is not clear if changes to the weighting strategy have been made over time.

In the 4th EQLS, Eurofound made a number of changes to the weighting procedures based on recommendations resulting from an external assessment of the weighting strategy following the 3rd EQLS (Vila & Cervera, 2014), For example, prior to the 4th EQLS, only weights for respondent selection at the household level were used, but more complex weights were introduced in the 4th wave. Other changes included the use of regression for the post-stratification adjustments, the addition of adjustments by household size, and the harmonisation of geo-location codes. We agree with the changes and overall approach to weighting in the 4th EQLS but changes to the weighting strategy may jeopardise comparability with past waves of the survey. The effect of weighting changes should be investigated to determine whether weights from previous waves of the EQLS should be recalibrated based on the new methodology.

5. Summary findings

Overall, we find that the design and implementation of the 4th EQLS largely reflects, and in some areas exceeds, current best practices for 3MC surveys. We also find that Eurofound has made important strides in advancing survey quality in the 4th EQLS through the innovative development and continued improvement of the QAP as a tool for assuring and controlling quality. In this section, we highlight the strengths of the 4th EQLS found throughout our assessment. This is followed by a discussion of some of the challenges or areas for improvement and some overall comments about the survey. The assessment concludes in Section 6 with a number of recommendations for continuous process improvement.

Perhaps the greatest strength of the EQLS is the extent to which the sampling frame development and many of the fieldwork processes were standardised across countries, including: the respondent selection process at the household level (with the exception of those few using individual registries), the use of a standardised CAPI instrument for both sample management and questionnaire administration, the mode of initial contact (with the exception of Sweden), the pilot test protocol, and number of call-backs. In 3MC surveys, while strict standardisation is neither always possible nor desired, organisations strive for at least some level of standardisation of processes and quality standards to facilitate comparability, as well as standardisation of associated monitoring and documentation (Pennell et al., 2017), This level of standardisation is largely possible due to the centralised nature of Eurofound and its role as a strong coordinating centre, which is crucial to maintaining adherence to quality requirements.

The questionnaire development process for the 4th EQLS incorporated many best practices including consultation with subject matter experts, a translatability assessment, team translation, and a sizable investment in pretesting. Pretesting for the 4th EQLS largely meets the minimum, and in some ways exceeds, the best practice we defined for pretesting in 3MC surveys.

Overall, changes in weighting between the 3rd and 4th wave were positive and more in line with best practices with regards to weighting. Key surveys weights were developed and well-documented, including guidance to data users on the correct use of weights for different types of analyses.

Comparisons of the sample composition of the 4th EQLS with data from Eurostat and the ESS suggest that the sampling and fieldwork processes are in line with other major cross-national data sources. Analysis of both the coefficient of variation and the design effects provide evidence of increased efficiency, which lessens the impact of subsequent weighting calculations on estimates of statistical precision, while increasing comparability. Examining response rate statistics, we also saw improvement in response rates and contact rates in a number of countries compared to the 3rd EQLS. The adoption of the AAPOR standard outcome codes for the 4th EQLS also facilitated the harmonisation of outcome codes and comparability to past EQLS waves and other 3MC surveys.

There is a growing awareness of the need for increased documentation and transparency in the field of 3MC survey research. However, no standard currently exists and documentation is particularly sparse in many 3MC surveys, which hinders efforts to assess comparability of data across surveys (Kolczynska & Schoene, 2018), While we see further areas for improvement, as noted in our recommendations, we recognise the substantial level of effort Eurofound has devoted and sizable amount of documentation produced related to key survey processes.

We next summarise the challenges and areas for improvement identified in our assessment. As noted in Section 4.1, Eurofound sets a minimum sample size for each country in the EQLS, with relatively little variation. While countries can opt to increase their sample size at their own cost, this is rarely done. As the EQLS uses a central funding structure, however, Eurofound is in a position to calculate the specific net sample required to achieve the desired effective sample size in each participating country, which could lead to a potentially more efficient use of resources. Further, the 4th EQLS relied on an enumeration method of sampling frame development for more than half the countries, which as discussed in Section 4.1 is more vulnerable to error than a list-based design.

With regards to the development of new questions for the 4th EQLS, it was difficult to assess the questionnaire without understanding what the analytical objectives were for the questions or how they would be used to inform specific analyses. Nonetheless, we identified some potential issues in our review of the final source questionnaire that may have been detected by an expert review carried out by a methodologist with training specific to questionnaire design in 3MC surveys. We have little doubt that the cognitive interviewing studies that were carried out added value to the development of the questionnaire, however we note ways that cognitive interviewing could be improved and used to validate question constructs and comparability in addition to identifying potential problems. Also, some new items added to the questionnaire resulted in significant item nonresponse due to screening questions, which was likely apparent in field pretests.[31] Ethical obligations require researchers to consider respondent burden when designing questionnaires, and items that will result in data difficult to use in analyses should be removed after such pretest results.

While fieldwork operations for the 4th EQLS were largely positive, we note some concerns with the timing of the fieldwork. Data collection for the 4th EQLS began in September, meaning that any delay in fieldwork would cause production to trail into the holiday season in December, a time when response rates may be impacted, human resources managing data are away from the office, and the interviewing staff may have less availability, to name a few negative effects, resulting in data collection to continue into January and beyond. Although preliminary analysis by Eurofound examining the relationship between substantive data and interview timing in the 4th EQLS found no emergent pattern, as December is the month when holidays are most likely to disrupt fieldwork procedures, it is advisable to begin fieldwork well ahead of the holiday season so as to complete data collection by mid-December (van Oostrum et al., 2017),

Response rates in the 4th EQLS increased in 12 countries and refusal rates went down or held constant in approximately half of the study countries. However, cooperation rates largely suffered in the 4th EQLS. These results underscore the importance of continued focus on contact rates to maintain or further improve response rates in future waves and also to examine potential nonresponse bias.

[31] For example, a univariate frequency distribution of the (unweighted) dataset indicates that only 1.5% of the entire EQLS sample provided a response to Q73a, *How satisfied or dissatisfied were you with the quality of the facilities (of long-term care services)?*

We also note that no partial interpenetration in fieldwork assignments was done which could allow for both the assessment and control of interviewer effects at the analysis stage. Further, paradata was not used optimally. Additional paradata should be collected and implemented in a responsive design framework both to increase the efficiency of the sample through increased interviewer monitoring while also increasing response rates.

Changes made to the weighting strategy introduced in the 4th EQLS could limit comparability with previous waves. The effect of the weighting changes needs to be investigated to determine whether resources should be directed to revising the weights from past waves to maintain the trend.

As noted above, documentation improvements could be made to further increase transparency and accessibility and to facilitate future quality assessments. For example, a lack of complete documentation on sampling frames, calibration statistics, and the target sample limits the assessment of coherence and comparability, both within the EQLS and to other surveys. We also encourage more detailed documentation of the development of the CAPI instrument, questionnaire development, the translatability assessment and the translation process.

In the future, it would be beneficial for Eurofound to involve a methodologist with specific training in survey methodology and experience in 3MC surveys at every stage of the project, from the development of the initial tender through the calculation and documentation of calibration weights and preparation of data for dissemination. Involvement could range from a full-time position, at one end, to temporary consultancies, at the other, but even periodic consultancies would prove advantageous to the quality of survey processes and outputs, and such expertise could be leveraged across all surveys at Eurofound.

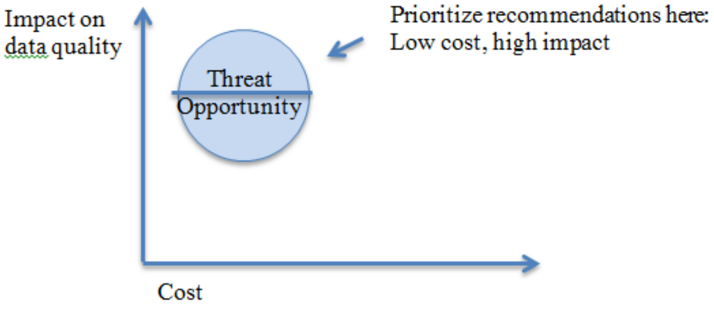
Based on a comprehensive review of the processes and outputs of the 4th EQLS, we find overall, that Eurofound has generally followed design, implementation, and quality control and quality assurance best practices for 3MC surveys, and in some cases exceeded the quality metrics observed in comparable surveys. We commend Eurofound for playing a leading role in advancing survey quality in the field of 3MC research.

6. Recommendations

We conclude the assessment with a series of recommendations, with prioritisation guided by the principles embodied in a SWOT analysis, which considers (s)trengths, (w)eaknesses, (o)pportunities, and (t)hreats, and is a framework to document ways to overcome threats and weaknesses and improve process and outcomes (Hill & Westbrook, 1997; Hofer & Schendel, 1978; Schnaars, 1998; McDonald, 1999; Kotler, 2000),

Opportunities and threats can be conceptualised as two sides of the same coin. That is, where there is a threat to survey quality (e.g., nonresponse bias), there is an associated opportunity to overcome the threat (e.g., adopting responsive design processes to minimise nonresponse bias and increase efficiency), The nexus of these concepts can be considered against both the cost of potential solutions as well as the subsequent impact on data quality. Here we draw on the strengths of the SWOT framework to prioritise recommendations by cost and impact, in each of the four main stages of the survey lifecycle, while simultaneously considering the impact in terms of the type of error addressed. Some recommendations are guided by the comparisons between best practices, ESS processes, and EQLS processes, while others are a result of considering a broader scope of design decisions faced by the EQLS and what opportunities for improvement might be considered in Eurofound's future surveys.

Figure 4. Recommendation Prioritisation Framework



Using this framework, we consider those items listed under the 'threat' and 'opportunity' sections of each task in the SWOT assessment and prioritise them using the framework illustrated in Figure 4, noting impacts on specific components within the TSE framework in the 3MC context. Tables 11 to 14, organised by survey lifecycle, list recommendations in order of priority, beginning with those we consider highest priority. We would identify as high priority recommendations in those areas where the threat and opportunity nexus is associated with a greater impact on data quality and associated error, but can potentially be addressed with a solution that is relatively lower cost. Areas where the threat/opportunity has a high impact on data quality but at a high cost, and where the threat/opportunity has a lower impact along with a medium or low cost, will be designated as medium priority. Considered within the framework of *fitness for intended use,* the majority of recommendations categorised as *high impact* speak to issues of accuracy. This is consistent with our discussion in Section 1.3 regarding the relative importance of this particular quality dimension.

We conclude this section with recommendations for Eurofound's reporting of output statistics, data dissemination, and data disclosure. Lastly, we provide recommendations for how the approach to quality, more generally, and how the QAP, more specifically, could be enhanced in the future.

6.1 Recommendations for sampling frame development

Table 11. Prioritisation of Recommendations for Sampling Frame Development Activities

	Low cost		High cost		
	High impact	Low impact	High impact	Low impact	Source of error addressed
Consider alternate respondent selection methods	X				Nonresponse error
Calculate effective sample size	X				Sampling error
Clearly define target the population		X			Coverage error, nonresponse error
Thoroughly document sampling frame sources		X			Comparison error
Consider alternatives to enumeration methods			X		Sampling error

Consider alternate methods of respondent selection at the household level to potentially reduce nonresponse.

While the process implemented in the 4th EQLS was standardised and followed best practices, concerns remain about the perceived intrusion of the full household roster and subsequent effect on survey cooperation. We recommend a hybrid approach where the interviewer first asks for the number of adults in the household (*n*), In households with one adult, the informant is selected. If there are two adults, the informant is sampled with a probability equal to *1/2*. In larger households (*n> 2* eligible persons), another method has to be used for respondent selection; preferably the Kish method (Rizzo et al., 2004; Koch, 2018), With fewer households subject to the full roster, it is possible that nonresponse will be minimised.

We conducted an analysis of the number of adults in households participating in the 4th EQLS, determining that the majority of countries had two or fewer adults, although estimates ranged widely from 92% (Denmark) to 50% in Albania, with sampled households in EU candidate countries having more adult members than in EU countries. Implementation of the method described above is low cost, but potential to reduce nonresponse error could be significant.

Require net sample size to vary across countries based on calculations of effective sample size necessary to achieve desired precision for each country.

Like all 3MC surveys, sample sizes in the EQLS are constrained by budget limitations. However, the centralised structure of the EQLS makes it possible for Eurofound to determine the net sample size required to achieve the desired effective sample sizes specific to participating countries, considering the design effects of each country. Eurofound should consider all aspects of the design, including response rates and design parameters (e.g., number of clusters, number of completes in a cluster, response rates by subgroup, etc.) in previous waves of its surveys as well as other, comparable 3MC surveys, to estimate the effective sample size (for step-by-step instructions see Heeringa & Ziniel, 2012; see also Kish, 1965), When resources are limited, such optimisation of resources can enhance comparability between countries and achieve comparable precision in a 3MC survey. Such calculations are low cost and the potential increase in precision provided could be significant.

The target population as defined in the project documentation should be consistent with the process used to determine respondent eligibility.

As noted in Section 4.1, the target population definition, as documented, may not have been implemented when assessing eligibility of potential respondents during fieldwork. It is important that the documentation accurately reflect the process followed by the interviewers. For example, if residence and language are important defining features of the target population, then the interviewer protocol and associated roster data should incorporate these criteria in the selection process. If the target population is not defined by these criteria in practice, then the documentation (e.g., the Technical & fieldwork report) of the target population should not include these criteria. Regardless of the specific target population definition, thorough documentation of eligibility screening processes is important for data users.

Sampling frame documentation should include information about the extent to which each individual or household register is up-to-date, as well as the date of data extraction.

Current documentation of the sampling frame development process in the 4th EQLS does not contain detailed information on specific sources and dates of access, both in terms of population/household registers, and in data used for geographic stratification. Without this data, users cannot assess whether two different surveys of a specific country are comparable (e.g., even if the same register was used in two countries, the recency of the specific data extracted cannot be determined), With minimal cost, thorough documentation leads to an increase in coherence and comparability for data users.

Enumeration methods are vulnerable to error and existing sampling frames should continue to be evaluated.

Researchers have expressed concerns about enumeration methods in sampling frame development, with evidence suggesting that these listing procedures are vulnerable to undercoverage error, particularly in specific types of housing unit segment characteristics, a particular concern when patterns of undercoverage are correlated with key variables of interest (Eckman & Kreuter, 2013; Bauer, 2016), Topics covered in the EQLS, such as access and satisfaction with public services, are likely to be correlated with housing units most prone to undercoverage. The potentially sizeable impact on sampling error leads to the recommendation that Eurofound continue to investigate availability of existing sampling frames with each round of surveys.

Currently, Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) is developing a series of work packages related to 3MC research across Europe. A recent SERISS publication provides a review of registers throughout Europe, noting availability and accessibility (Scherpenzeel et al., 2017), Although Eurofound is not a member of this consortium, it would be advantageous to consider potentials for partnering with other organisations undertaking 3MC research to leverage economies of scale and develop sampling frames in countries without registers. In the shorter term, Eurofound should review the most recent publically available deliverables in survey development available from SERISS.[32]

Lastly, particularly if collaboration with other organisations is not possible, Eurofound could consider alternate register sources such as utility companies in those countries without a more obvious available register, or consider a full list of housing units in selected PSUs prior to household selection, rather than selecting and listing only specific households during the enumeration process.[33] Although such solutions are associated with increased cost, the potential for error reduction could be substantial.

6.2 Recommendations for questionnaire development activities

Table 12. Prioritisation of Recommendations for Questionnaire Development Activities

	Low cost		High cost		
	High impact	Low impact	High impact	Low impact	Source of error
Develop research questions/aims and an analysis plan for new questions	X				Specification error
Involve a survey methodologist trained in 3MC questionnaire design	X				Specification error, measurement error
Carry out an expert review	X				Comparison error, measurement error
Adopt a uniform approach to assessing final translations	X				Measurement error
Produce show cards and standard protocol for their use	X				Measurement error
Expand documentation of the questionnaire development process		X			Comparison error

[32] See https://seriss.eu/resources/deliverables/.

[33] For an example of this latter recommendation, see the discussion on the process used to construct the sampling frame in Lebanon for the World Mental Health Survey (Heeringa et al., 2008),

Consider using advance translation or a team approach to the translatability assessment			X		Measurement error
Expand the use of cognitive interviewing			X		Measurement error

Clearly outline research questions/aims and an analysis plan for new questions.

Research questions/aims and an analysis plan should be developed for new questions added to the EQLS going forward. Clearly defining the research questions or aims and developing an analysis plan is important for guiding the question development process and helps ensure that all necessary questions are included in the questionnaire and that unnecessary questions are avoided. This is important from a practical standpoint in making the interview as efficient as possible as well as from an ethical standpoint in reducing undue respondent burden.

Involve trained survey methodologists in 3MC questionnaire design at key stages of the questionnaire design process.

It would be beneficial for Eurofound to involve a survey methodologist at key stages of the questionnaire design process. Subject matter experts are experts in a particular field but often do not have expertise in the design of survey questions. Therefore, it is valuable to include a survey methodologist in consultations with subject matter experts. This allows the subject matter experts to focus on the content of questions, while the survey methodologist can advise on the best ways to go about asking the questions. Further, a survey methodologist or ideally more than one should also carry out an expert review of the questionnaire, as discussed in more detail below. Survey experts could be brought in as consultants at key stages of the questionnaire design process, which would involve relatively low cost. However, it may be worthwhile for Eurofound to invest in adding a full-time survey methodologist who could advise on all aspects of survey design and implementation for the EQLS as well as the other surveys in Eurofound's portfolio.

Carry out an expert review of the source questionnaire.

Expert review consists of a review of a draft questionnaire by a small group experienced methodologists or subject matter experts (usually 2-3), Subject matter experts evaluate the whether the content of the questions is appropriate for measuring the intended concepts and meets the analytic objectives of the survey. Questionnaire design experts assess aspects of the questionnaire including question wording, structure, order, response alternatives, interviewer instructions, and skips patterns. For an overview, see Groves et al. (2009), Expert reviews offer a relatively low-cost and effective method for evaluating questions and identifying potential problems, particularly when used in combination with cognitive interviewing (Maitland & Presser, 2017),

Adopt a uniform approach to assessing final translations.

As noted in Section 4.2, it is not clear that a final uniform translation quality assessment step took place. It is important that all translations are subjected to the same reviewing steps to ensure a homogeneous level of translation quality. We recommend the addition of a translation quality assessment step as part of the translation process. This step should be carried out on pre-final translations before pretesting so the outcomes can be incorporated when finalizing the translations before the instruments are fielded. As practiced by the ESS, translation quality assessment could consist of (a) translation verification by cApStAn, that is, a linguistic assessment striving for semantic and pragmatic similarity to the source text, and (b) formal checking of the translations using the Survey Quality Predictor (SQP), which examines similarity between the translated questionnaires and the source text based on a set of formal criteria.

Produce show cards for survey items as needed and a standard protocol for their use.

As noted in Section 4.2, separate show cards were not used in the 4th EQLS. Instead, interviewers were instructed to show response categories to the respondents on the tablet screens for particular questions, as indicated in the Source Questionnaire. While literature on the use of show cards in 3MC research is scarce, they are used in comparable surveys such as the ESS and SHARE. We highly recommend that show cards be produced for survey items as needed and that a standard protocol be developed for their use by interviewers in all countries.

Expand documentation of the questionnaire development process.

We are not aware of detailed documentation related to the development of new EQLS questions or to modifications made to questions at different stages and why. We recommend the use of a spreadsheet so that it is possible to see changes to questions over time and the reasons why changes were made. For example, the response options to a question may have been modified due to little variation in responses in past waves. It could also be noted if modifications to the questions were reflected in any/all translations as well. It would be helpful to include the sources for questions that may have been adopted from other surveys.[34]

As part of the questionnaire development process, we also recommend more detailed documentation of the translatability assessment and the translation process. For the translatability assessment, we recommend that future documentation include information on the skills and qualifications of those involved, the instructions or training provided, the outcome of the comments (e.g., changes in the source questionnaire, notes for the translating teams, etc.), a synthesis of the comments from all languages and information about how and why (or why not) the TA comments were included in the finalisation process of the source questionnaire, and how the TA process was organised by the contractor or how queries from the translators and/or the contractor towards Eurofound were handled.

For the translation process, we recommend that the documentation be more comprehensive, and that the terminology used conform to the terms used in the classic TRAPD approach (e.g., ‘adaptation’, ‘harmonisation’, ‘third adjudicator’, etc.) (see Dorer, 2015),

[34] Eurofound does maintain an in-house Question compendium and information about particular survey items can be requested by contacting Eurofound staff.

Consider carrying out advance translation or adopting a team approach for the translatability assessment (TA) and allow results from a TA or AT to inform revisions to the source questionnaire.

We recommend that Eurofound adopt a team approach for the translatability assessment or consider using advance translation, which involves a team approach. In a team approach, as discussed above, interdisciplinary teams composed of translators and survey researchers apply a multi-step translation approach and discuss both the translators' and the survey researchers' comments in a review session. This helps ensure that the final comments include the input of someone trained in survey research and are as relevant as possible for the translation of the source questionnaire into the various target languages. Because a team of at least three experts per language is typically needed for a team approach, it would involve higher cost than the typical TA exercise. However, we believe that the increased quality and relevance of the results could make a sizable impact.

We also recommend that the AT or TA be carried out earlier in the questionnaire development and translation process so that the results can feed into possible modifications to the source questionnaire. This would enable Eurofound to take more complete advantage of the benefits of a TA or AT, which is to help make the source questionnaire as easy to translate as possible into the target languages.

Expand the use of cognitive interviewing to validate question constructs and comparability and with additional language families, major regional subgroups, or countries as resources permit.

As noted above, results from a recent study suggest that the best combination of question evaluation methods may be expert reviews followed by cognitive interviews (Maitland & Presser, 2017), We therefore strongly support Eurofound's continued use of cognitive interviewing as part of the questionnaire development process. However, as noted above, in addition to a method to identify potential problems, cognitive interviewing can also be used to assess construct validity and comparability, which offers a particular advantage and opportunity for cognitive interviewing in the 3MC context. Miller (2018) describes in detail the analytic goal of comparative cognitive interviewing studies, the relevant components of such a study, including data collection (e.g., the structure of the interview and data quality), data analysis techniques, and strategies and tools for conducting such studies.

We recommend a comparative cognitive interviewing study carried out in a core sample of countries similar to that used for the TA for the 4th EQLS that covers the major language groups and geographic regions of EU countries. However, the choice of countries could also be informed by where local experience and expertise in cognitive interviewing is available. Online probing (Meitinger & Behr, 2016; Behr et al., 2014) and crowdsourced cognitive interviewing (Murphy et al., 2013) are recent approaches that could offer a cost-effective way of confirming results found in an initial round of face-to-face interviews as well as potentially cover additional language and national subgroups of respondents (see Braun et al., 2015 for discussion),

6.3 Recommendations for fieldwork

Table 13. Prioritisation of Recommendations for Fieldwork Activities

	Low cost		High cost		
	High impact	Low impact	High impact	Low impact	Source of error
Thoroughly document the CAPI development & testing process	X				Measurement error
Conduct nonresponse bias analyses	X				Nonresponse error
Standardise the mode of initial contact	X				Nonresponse error, coverage error
Implement partially interpenetrated fieldwork assignments and conduct interviewer effects analysis			X		Measurement error. nonresponse error
Implement responsive design techniques			X		Nonresponse error
Consider alternate interviewer pay structure			X		Measurement error, nonresponse error

The CAPI instrument development and testing process should be thoroughly documented.

The implementation of a standardised CAPI instrument is significant in reduction of several sources of survey error. However, while testing is a critical component of the technical development process, documentation of procedures for quality assurance to ensure consistency in the instrument itself is absent in the 4th EQLS. We recommend that the central coordinating centre provide a clear set of instrument specifications and/or a data dictionary for the instrument, which will facilitate testing and assessment of the CAPI instrument. Such documentation would include: question (variable) names and labels; question text; response option values and labels; numeric response formats and ranges, and specifications for the lengths allowed for open-ended question text; interviewer or respondent instructions; missing data values; skip instructions; and so on. It should enable comparison of computerised or formatted paper instruments to instrument design specifications. We also recommend centralised instrument evaluation procedures, such as expert review against heuristics, review of instruments, data dictionary, or codebook to assess adherence to instrument specifications, and usability tests.[35] After testing is complete, there should be

[35] See specifically Guideline 5 in Hansen et al., (2016) for a more extensive list of testing options.

subsequent visual documentation of the instrument itself to facilitate assessment. Solutions for documentation need not be high-tech, and simple screen shots of each screen in the instrument would be relevant for a comprehensive review of the technical design. While documentation costs are low, cost for testing recommendations can vary but even lower cost options could have significant impact in reduction of error.

Partially-interpenetrated fieldwork assignments should be implemented to facilitate subsequent in-depth analysis of interviewer effects.

As discussed in Section 4.3, geography and interviewer assignment are at least partially confounded in the data collected in the 4th EQLS. While a full interpenetrated design – where interviewers are assigned at random completely across all PSUs – is nearly always cost prohibitive, a partially interpenetrated design still results in data which can be used to investigate the relative magnitude of the variance explained by specific interviewer characteristics. For example, one analysis showing a strong association between interviewer attitudes and respondent attitudes, and another showing evidence of variance in interviewer behaviour with regards to requesting a private setting in an interview, has led to consideration of those elements in interviewer training that could be emphasised so that increased standardisation in interviewer behaviour can be achieved (Mneimneh et al., 2017; de Jong et al., 2017), Such analyses can have direct impacts on informing measures to minimise measurement error in future surveys. There is additional cost involved in implementing interviewer assignment protocols due to increased interviewer travel and time, but Eurofound may consider selecting a sample of countries – perhaps one from each region – where this interviewer assignment is implemented, allowing for an initial cross-national analysis of interviewer variance and potential effects on substantive results.[36]

Additionally, while contributing to survey methodology is not Eurofound’s primary focus, we note that literature on interviewer effects in 3MC surveys using a partially interpenetrated fieldwork assignment is scarce, and indeed the possibility of interviewer effects in analysis is rarely considered (Beullens & Loosveldt, 2016; Elliott & West, 2015), Implementing a design in the EQLS would make the data more attractive to methodologists as well as researchers in other disciplines who are interested in the way in which methodology may impact their findings, leading to potentially increased visibility and utility of EQLS surveys.

Conduct nonresponse bias analyses after fieldwork is complete.

Eurofound should consider performing a basic response analysis for all participating countries and establishing a threshold for extensive nonresponse bias analysis in order to prioritise resources (PIAAC, 2010),[37] In order to conduct a nonresponse analysis, data must be available for all sample units and be collected and coded consistently across all cases and in all participating countries (Blom et al., 2008), In the 4th EQLS, data collected for non-interview cases was limited to final outcome code, timing of (attempted) contacts, and reasons reported by nonresponse households/individuals, as recorded by the

[36] For a concise discussion of interpenetrated assignment, see Groves et al. (2009), pp. 296-297.

[37] The protocol for nonresponse analysis detailed in PIAAC (2010) is comprehensive and may be an appropriate framework for Eurofound to consider as a basis in the future.

interviewer (e.g., *I'm too busy¸ Survey takes too long*, *No motivation*, etc.), The list of potential outcome codes was also expanded in the 4th EQLS and provides additional opportunity to study nonresponse. Requiring interviewers to obtain additional observational data would provide more comprehensive information about non-respondents and how they might differ from respondents. Such data might include information about the home and neighbourhood of each sampled unit, as collected in the ESS, where previous analyses have provided evidence that data on neighbourhood characteristics may allow for correction of nonresponse bias (Stoop et al., 2010), Observation data is vulnerable to error, and interviewer training should include a section focused on the collection of these data, with a component where interviewers practice their skills and inter-rater reliability is assessed (Campanelli et al., 1997; Sinibaldi et al., 2013; West & Kreuter, 2013),

Evidence from a meta-analysis suggests that bias in demographic variables is not predictive of the difference in substantive variables between respondents and non-respondents (Peytcheva & Groves, 2009), This indicates an increased importance in identifying and recording additional data from non-respondents thought to be correlated with key variables of interest in the EQLS. However, such additional (observational) data collection is relatively low cost and has the potential for significant impact for reducing nonresponse in the future among specific subgroups if information can be obtained about non-respondents and used to target them to increase response rates. Subsequent nonresponse bias analyses can also inform the quality of interpretation of data after fieldwork.

Standardise the mode of initial contact among countries.

Interviewers' initial contact with sampled households/respondents was face-to-face in all participating countries in the 4th EQLS, with the exception of Sweden. In Sweden, a register of individuals was used as a sampling frame, and the Technical and Fieldwork Report for the 4th EQLS notes that due to the sparsely populated regions of the country, respondents are generally pre-recruited respondents by telephone. Prior to fieldwork, individuals selected from the register in Sweden were matched with telephone numbers from the register and other sources. However, as illustrated in Table A6 in Appendix 3, the contact rate for Sweden was only 64%, with a corresponding response rate of 16%, with both rates having decreased significantly from those reported for the 3rd EQLS. This suggests that use of the telephone as the initial mode of contact may not be as effective as even a few years ago, as people may be less likely to answer phone calls from unknown numbers. Based on the outcome in Sweden, we recommend that Eurofound require initial contacts in all participating countries to be face-to-face, regardless of typical practice. While Sweden is more sparsely populated than some countries in the survey, it is not unique and because a cluster design is implemented, the increased cost may not be that significant but could lead to a substantial increase in quality if resultant nonresponse rates decrease.

Implement a responsive design approach.

Some organisations have used techniques such as responsive (or adaptive) survey design, to monitor survey data collection processes and adapt study designs in real time (Groves & Heeringa, 2006), The fieldwork process in the 4th EQLS introduced an element of responsive design through the application of sample replicates, wherein an initial sample of cases was released in each PSU, and then response rates were monitored to determine the extent to which subsequent replicates should be released.[38] Section 3.3 discusses analyses which indicate that the replicate release process led to increased efficiency. When implemented accurately, as in the 4th EQLS, a replicate release design presents no threat to probability of selection (Kalton, 1983; Lohr, 2009), However, there are a number of other elements of responsive design whose application can also significantly impact both efficiency and the overall quality of the data (Groves & Heeringa, 2006; Marker & Morganstein, 2004; Pierchala & Surti, 2009; Thompson & Oliver, 2012),

Dashboards – visual displays of information on costs, timeliness, and quality across processes – can be used to monitor field progress, looking specifically at measures of inputs (hours, sample), output (interviews), efficiency (costs per interview), GPS data, and quality (response rate, changes in key estimates, key stroke data), and make adjustments to field processes as necessary.[39] This assessment of the 4th EQLS has suggests that the EQLS would benefit from implementing tighter field control procedures in some areas (e.g., deviations in assignment of final outcome attempts noted in Section 2.3), as well as evaluation and verification of not only those interviews which were completed, but also those cases which did not result in a completed interview.

Dashboard analyses can also inform both more efficient release of replicates, as well implementation of a two-stage responsive design, wherein a subsample of non-respondents from the initial sample is selected for an intensive contact protocol and/or refusal conversion in the second phase.[40] Such an approach can truncate the increase in costs near the end of the field period by focusing interviewers on these selected cases. We recommend that Eurofound consider selecting a subset of countries, e.g., one/two countries in each region of Europe; or one/two countries each having low/medium/high relative nonresponse in the 4th EQLS to implement a more comprehensive monitoring system as well a two-stage responsive design to target nonresponse and more closely examine nonresponse bias in the second stage. Results found to decrease nonresponse bias and end of survey costs would provide evidence for further development of these approaches.

As discussed in Section 3.3, changes in sample implementation likely contributed to the decrease in the coefficient of variation and the design effects across the majority of countries. However, the impact of this change in the design feature is captured by the design weights and, providing weights are applied correctly in analysis, does not impact either quality or comparability of data from the 4th EQLS with that of previous waves. Likewise, any changes made through implementation of protocols related to responsive design will not impact the ability to conduct comparative analyses between data from the 4th EQLS and future waves providing that appropriate documentation occurs, and relevant weights are calculated.

[38] In the 4th EQLS documentation, this process was referred to as ‘batch release’. Adopting the terminology most commonly used in survey methodology literature, in this assessment we refer to this process as ‘replicate release’.

[39] For a discussion and detailed examples of using paradata to monitor fieldwork, see Kirgis & Lepkowski (2013), Mneimneh et al. (2018), and Hyder et al. (2017), For a visual representation of *dashboards*, see Biemer (2010),

[40] For a detailed discussion of this use of responsive design, see Groves et al., 2009 (Chapter 6), Wagner et al. (2012); Groves & Heeringa (2006); Axinn et al. (2011), For a discussion of use of paradata and other analyses pertinent to detecting data fabrication, see Robbins (2018),

Complement strategies used in responsive design with deliberate decisions about interviewer pay structure.

Implementation of many of the strategies stemming from the responsive design framework are reliant upon effective interviewers. The structure for interviewer pay (e.g., per interview, hourly, salary, etc.), differs by data collection firm and is often based in the country's research tradition. Levels of interviewer pay, as well as the pay structure, can affect interviewers' motivation to implement all aspects of the research protocol. If an interviewer is paid for the number of hours actually worked (*e.g.*, hourly wage or weekly salary), s/he is more likely to devote time to other important aspects of data collection, such as refusal avoidance. Payment on a piecework basis increases the risk that the quality of an interviewer's work will suffer. Documentation on interviewer pay structure in the 4th EQLS is scarce, but a review of ESS protocols indicates that interviewers are most commonly paid per interview, and as there is partial overlap in data collection firms used in the most recent rounds of the EQLS and ESS, this suggests a fee-for-interview pay structure is most prevalent in the 4th EQLS as well.

We recommend that in future surveys, Eurofound considers following the model implemented in PIAAC, where the standard clearly states 'The basis for remunerating interviewers for their work must be independent of the number of completed interviews. In other words, interviewers are not to be remunerated on a piecework basis' (PIAAC, 2014, p. 119), The standards also include protocols regarding pay rates and incentives.[41]

6.4 Recommendations for weighting

Table 14. Prioritisation of Recommendations for Weighting Activities

	Low cost		High cost		
	High impact	Low impact	High impact	Low impact	Source of error
Investigate the effect of weighting changes	X				Comparison error, adjustment error
Provide control data and information about the sources		X			Comparison error
Document how missing data was handled		X			Comparison error

Investigate the effect of weighting changes introduced in the 4th EQLS.

[41] See PIAAC (2014), Guideline 8.3.5 for further details on PIAAC's protocol for interviewer pay.

We strongly recommend investigating the effect of the changes in weighting on comparability between the 4th and 3rd EQLS to determine whether the weights for previous waves should be recomputed based on the revised weighting strategy. Such a study would involve generating new weights for the 3rd EQLS based on the methods used for the 4th EQLS for a selection of countries and variables. Estimates based on the previous and revised weights could then be compared and the impact of the changes in weights assessed. This exercise should also be carried out going back further than the 3rd EQLS if comparison over time past the 3rd EQLS is a high priority. Recalibrating weights for previous EQLS waves based on the new weighting strategy would be ideal for maintaining comparability over time. While this could be a challenging undertaking in terms of locating appropriate past external data sources for every country and costly in terms of human resources, it should be considered if the impact of the revised weights is substantial and if comparability going back is a chief goal. Investigating the impact of the revised weights could help inform decisions about whether and the extent to which weights should be revised.

Provide the control data used for post-stratification weights and more information about the sources.

Documentation and available data for the 4th EQLS does not include the control data used for post-stratification or detailed information on specific sources and dates of access. Without this data and information, it is not possible for users to closely examine the data used for post-stratification, compare it with that used for other countries or other surveys, or to replicate or compute different weights. The data and information provided by the ESS in this regard could serve as a guide.

Documentation should be provided related to how missing data is handled in weighting.

Overall, documentation about weighting in the 4th EQLS is sufficiently detailed with the exception of information about missing data for the variables used in weighting. We suggest including information about the amount of missing data and how it was handled.

6.5 Recommendations for web add-on

With increasing although still uneven internet penetration across the EU, alongside the increasing costs of both face-to-face interviews and increasing nonresponse, interest in online surveys has become more prevalent. In Germany, Poland, Slovenia and the United Kingdom, the 4th EQLS tested a web survey as a new component (referred to as a ‘web add-on’) following face-to-face data collection. According to Eurofound, the stated purpose of the exercise was to investigate mode effects among those items in the questionnaire perceived to be more sensitive, to investigate the feasibility of a follow-up web survey, to learn more information about those cases who were non-respondents in the face-to-face survey, and to include additional questionnaire content not included in the face-to-face interview. Both respondents and non-respondents in the face-to-face survey were targeted for participation. Those in the former group were invited to provide an email address at the end of the interview to receive a subsequent email invitation or received a password from the interviewer to complete the web survey online. Those in the latter group received an explanatory letter and password, provided either in person or left in the mailbox.

Resulting response rates were relatively low across all countries, particularly among those who were non-respondents in the face-to-face survey.[42] In this section, we briefly review the extent to which the documentation indicates compliance with the practices outlined in Section 4 as relevant to the web add-on (i.e., questionnaire development and fieldwork), consider similar activities in comparable studies, and offer recommendations for supplementary use of web surveys in the future.

The 4th EQLS web add-on questionnaire included both the repetition of some questions from the face-to-face survey as well as a number of new items. Available documentation indicated the following: 1) interviewer training included a standardised protocol on introducing the web survey to respondents as well as inviting non-respondents to participate; 2) a pretest of web add-on protocols occurred alongside the face-to-face pretest; and 3) incentives were used and documented. While available documentation indicates that web add-on items were subject to the same translation process as the face-to-face items, with the exception of adjudication, there is no evidence that these new items were subject to cognitive testing by Kantar Public or to any initial development work conducted by Ipsos. Documentation of the technical instrument development and design is also missing, as is any reference to the use of paradata to measure respondent behaviour in the web survey.

The ESS has conducted several mixed mode studies since 2003, with each study aimed at a specific research question and fielded in a small number of countries. Overall findings suggested that the face-to-face design performed better than any alternative web surveys that were tested in terms of survey participation and net sample representativeness (Villar & Fitzgerald, 2017), To meet this challenge, the ESS has recruited a probability sample from participants from three countries (Estonia, UK, and Slovenia) in ESS Round 8 to develop a cross-national online panel, providing those respondents lacking internet access with an Internet-enabled tablet and an incentive. Preliminary findings suggested response rates ranging from 49% in the UK to 69% in Estonia (Villar et al., 2017), Initial agreement rates were relatively high, although actual participation differed across countries, and those without Internet were more difficult to recruit.

Specific recommendations for improving data quality from web add-on surveys differ depending on the primary objectives of Eurofound. For example, if the focus is testing mode effects vis-a-vis sensitive questions, then expending significant resources to increase response rates may not be an optimal use of resources. However, if the objective is to use the platform to collect additional data from face-to-face respondents, then Eurofound should consider a study design where a technical device with means to access the Internet is provided to those without Internet access, which has a significant associated cost. An overall recommendation, then, is that any future use of web surveys be limited to addressing one specific research question in order to optimise resources and collect high quality data.

Additionally, based on both the challenges encountered and industry best practices, we recommend that Eurofound consider the following if administering similar web-based surveys in the future:

- All survey items specifically for inclusion in the web add-on questionnaire should be developed completely alongside face-to-face items, including cognitive testing, aforementioned expert

[42] See Kantar Public's Web Survey Technical and Fieldwork Report for further detail on response rates.

review by a survey methodologist, and all steps of the translation process, to maintain quality standards.

- There should be comprehensive documentation of the technical platform used in the web add-on. The visual appearance of a web survey can impact response, particularly the extent to which the survey is optimised for a desktop/laptop or mobile device. As there is no available documentation, we cannot assess whether the visual appearance may have impacted response. Recommendations noted in Section 6.3 regarding CAPI development and testing apply here as well.
- Online survey platforms can retain a variety of paradata, depending on the specific respondent, and such paradata can provide information about how respondents interact with the web survey. For example, variance in device usage both across respondents and at the individual level (i.e., device switching) can impact response, survey breakoff and timing of survey completion. An increased understanding of device usage and device switching through examination of answers to these questions can inform future efforts to minimise nonresponse in the EQLS (Cheung, 2017),
- The most significant irregularities (incorrect protocol followed regarding web add-on invitation and data linking) occur within the scope of the interviewers' activities. Kantar Public's documentation references the increased interviewer burden due to differences in whether respondents provided an email address, and due to the fact that both respondents and non-respondents were targeted. As discussed further below, refinement of web add-on objectives should result in redefinition of the web add-on target population and ideally a reduction in subsequent interviewer burden and associated error.
- In addition to other suggested revisions (see Section 6.7), web-specific indicators should be added to the QAP.

6.6 Recommendations for data and results dissemination

After completion of each round of the EQLS, Eurofound has produced a comprehensive report to share the results of the survey, based primarily on univariate distributions and/or averages of individual survey items, as well as indices calculated from specific series of items.[43] These data are generally presented at the country-level in league tables, which are often used in 3MC surveys and which encourages comparison of country level data.[44]

Most often, these league tables do not provide users with the ability to assess whether differences between countries are statistically significant. This is a critical issue because without these explanatory statistics, media and policy makers may draw possibly erroneous conclusions both about individual countries as well as similarities and differences between countries (Yasukawa et al., 2017; Lyberg, 2016), Therefore, we recommend that Eurofound consider reporting confidence intervals in league tables, displaying a point

[43] See Eurofound (2017) for the EQLS 2016 Overview report.

[44] For example, the ESS publishes the *ESS Topline Series*, a set of comparative cross-national analysis based on specific topics in the questionnaire. See http://www.europeansocialsurvey.org/findings/topline.html for the most up-to-date series.

estimate for each country, with a shaded bar from lower confidence limit to upper confidence limit. Such a visual display will give readers a sense of the extent of uncertainty around these statistics.[45]

In addition to publishing analyses using the data, Eurofound also provides the data for each wave of the EQLS in a public-use dataset. Data is accessible via the UK Data Service, which carries out checks prior to making the dataset accessible to general public. Data dissemination and archiving requires careful consideration of issues related to non-disclosure and respondent confidentiality, particularly as the new General Data Protection Regulation becomes effective in May 2018. When preparing data for public release, the questionnaire and dataset should be inspected for *indirect identifiers*, key identifying sociodemographic characteristics in the data which make unique cases visible and possibly identify a respondent. There are a variety of methods to treat such data before inclusion in a public-release dataset, including removal, top-coding, collapsing, among others.[46]

Geographical information should never identify the PSU. Eurofound may also consider the use of pseudo-PSUs, which mask a respondent's specific PSU but retain the information regarding to which stratum a PSU belongs (Heeringa et al., 2010),

6.7 Recommendations for quality assurance

We provided an assessment of the extent to which the quality indicators were achieved in the QAP in Section 2.5, noting its utility as one component in a multi-faceted approach to assessing survey quality, alongside the other approaches advanced in this assessment. However, we recommend several revisions to improve the efficacy of the QAP in future surveys.

- Current quality indicators should be revised to ensure that they are consistently singular, measurable, and unambiguously specified, as noted in the examples of challenges we provide in Section 2.5.
- We recommend removing the designation of 'ideal world' from any indicators in the QAP beyond the tender process. In general, failure to meet the required targets, which typically represent a minimum standard, could be considered to have the greatest relative impact. However, the designation of real-world and ideal-world does not clearly imply that one is more important than the other.
- New quality indicators, inspired by best practices as outlined in this assessment as well as in the literature reviewed in our evaluation, should be added to the QAP. For example, an indicator guided by the best practice about fieldwork assignment might be 'Percentage of PSUs where all interviews were conducted by only one interviewer', with an associated target of 0%.

[45] However, the figures in Eurofound (2017) do provide information on whether the change over time was statistically significant, and the background checks regarding differences between countries or social groups have been carried out ('attention is drawn in the text only to differences or findings that are statistically significant (at 0.05 level)', p.9),

[46] The Inter-university Consortium for Political and Social Research (ICPSR), a leading data archive located at the University of Michigan, provides a comprehensive guide to preparing data for public release: https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/

Lastly, we offer a recommendation for a somewhat structural revision to the QAP. As discussed in Section 1.3, our assessment based on the QAP focused on accuracy related indicators, the achievement of which we found to have the most bearing for assessing overall survey quality. Indicators related to the quality dimensions of punctuality and accessibility had less utility in this regard. Section 2.1 includes a discussion of the importance of punctuality in terms of the sequential nature of certain tasks. Specific dates are less important to questions of quality. Similarly, the extent to which documentation is available – many indicators related to the quality dimension of accessibility address – do not affect the quality of the data itself.

The QAP is currently very lengthy, with about 145 indicators in total: 34 relating to accessibility, 31 to punctuality, and 73 to accuracy, while the final few relate to coherence and comparability and relevance and timeliness. Both to address the unwieldly nature of the QAP as well as to improve its utility by retaining only those indicators directly associated with data quality, we recommend removing all accessibility indicators and all punctuality indicators with the exception of those few related to task sequencing. Accessibility indicators should be revised into a list of documentation, further divided as relevant and distributed for use as a documentation checklist by Eurofound, the central coordinating centre, and participating countries. Punctuality indicators should be transformed into a study management tool to monitor work flow. The resulting QAP, consisting primarily of indicators relating to accuracy, would provide a more effective tool for quality assurance.

References

All Eurofound publications are available at www.eurofound.europa.eu

AAPOR (2016), Standards and definitions. Final dispositions of case codes and outcome rates for surveys. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf. Retrieved on March 19, 2018.

Ahrendt, D., Sandor, E., Burnett, J., & Gyuzalyan, H. (2017), The 4th European Quality of Life Survey: New elements introduced in the sampling strategy: Challenges and lessons learnt. Paper presented at the European Survey Research Association Bi-annual Meetings, July 17 – 21, Lisbon.

Axinn, W., Link, C., & Groves, R. (2011), Responsive survey design, demographic data collection, and models of demographic behavior. *Demography, 48*(3), 1127–1149.

Bauer, J. J. (2016), Biases in random route surveys. *Journal of survey statistics and methodology*, *4*(2), 263–287.

Behr, D., M. Braun, L. Kaczmirek, & W. Bandilla (2014), Item comparability in cross-national surveys: Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & quantity 48*: 127–148.

Benstead, L.J. (2014), 'Does interviewer religious dress affect survey responses? Evidence from Morocco, *Politics and religion, 7*, 734–760.

Benstead, L.J., & D. Malouche. (2015), 'Interviewer religiosity and polling in transitional Tunisia,' Paper prepared for the Midwest Political Science Association annual meeting, April 16-19, Chicago, IL.

Beullens, K. & Loosveldt, G. (2014), Interviewer effects on latent constructs in survey research. *Journal of survey statistics and methodology*, *2*(4), 433–458.

Beullens, K., Loosveldt, G., Denies, K., & Vandenplas, C. (2014a), Quality report for the European Social Survey, Round 6. London: European Social Survey ERIC.

Beullens, K., Matsuo, H., Loosveldt, G., & Vandenplas, C. (2014b), Quality matrix for the European Social Survey, Round 7. London: European Social Survey ERIC.

Beullens, K., & Loosveldt, G. (2016), Interviewer effects in the European Social Survey. *Survey Research Methods* 10:2, pp. 103–118),

Biemer, P. P. (2010), Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

Biemer, P. P., & Lyberg, L. E. (2003), *Introduction to survey quality*. John Wiley & Sons, Inc.

Biemer, P., Trewin, D., Bergdahl, H., & Japec, L. (2014), A system for managing the quality of official statistics. *Journal of Official Statistics, 30*(3), 381-415.

Billiet, J., Philippens, M., Fitzgerald, R., & Stoop, I. (2007), Estimation of nonresponse bias in the European Social Survey: Using information from reluctant respondents. *Journal of official statistics*, *23*(2), 135.

Blom, A. G., Lynn, P., & Jäckle, A. (2008), *Understanding cross-national differences in unit nonresponse: the role of contact data* (No. 2008-01), ISER Working Paper Series.

Blom, A., De Leeuw. E.D., & Hox, J. (2011), Interviewer effects on nonresponse in the European Social Survey. *Journal of official statistics, 27*(2), 359–377

Braun, M., D. Behr, L. Kaczmirek, & W. Bandilla (2015), Evaluating cross-national item equivalence with probing questions in web surveys. In *Improving survey methods: Lessons from recent research*, eds.

U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis, 184–200. New York: Routledge, European Association of Methodology.

Brick, J. M., and Williams, D. (2013), Explaining rising nonresponse rates in cross- sectional surveys. *Annals of the American academy of political and social science*, *645*(1), 36–59.

Campanelli, P., Sturgis, P., & Purdon, S. (1997), *Can you hear me knocking? and investigation into the impact of interviewers on survey response rates*. National Centre for Social Research.

Cheung, G. (2017), Device switching: What we learned from web survey logins. Paper presented at the International Workshop on Comparative Survey Design and Implementation, Mannheim, March 16-18, 2017.

Caspar, R., Peytcheva, E., Yan, T., Lee, S., Liu, M. & Hu, M. (2016), Pretesting. *Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved May 4, 2018, from http://ccsg.isr.umich.edu/pretesting.cfm

Cibelli Hibben, K., de Jong, J., Hu, M., Durrow, J., & Guyer, H. (2016), Study design and organizational structure. *Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Conway, K., Acquadro, C., & Patrick, D. L. (2014), Usefulness of translatability assessment: results from a retrospective study. *Quality of life research: An international journal of quality of life aspects of treatment, care and rehabilitation*, *23*(4), 1199–1210.

Couper, M. P. (1998), Measuring survey quality in a CASIC environment. Paper presented at the Proceedings of American Statistical Association: Survey Research Methods Section.

de Jong, J., Mneimneh, Z., &; Moaddel, M. (2017), Measuring third-party presence during face-to-face interviews: Respondent and interviewer predictors and effect on reporting sensitive attitudes in Jordan and Turkey. Paper presented at the International Workshop on Comparative Survey Design and Implementation, Mannheim, March 16-18.

Dimitrova, D. D., & Dzhambov, A. M. (2017), Perceived access to recreational/green areas as an effect modifier of the relationship between health and neighbourhood noise/air quality: Results from the 3rd European Quality of Life Survey (EQLS, 2011–2012), *Urban forestry & urban greening*, *23*, 54–60.

Dorer, B. (2011), Advance translation in the 5th Round of the European Social Survey (ESS), FORS Working Paper Series, 2011–2014, Lausanne: FORS.

Dorer, B. (2012), Round 6 translation guidelines. Mannheim: European Social Survey, GESIS.

Dorer, B. (2015), Carrying out 'advance translations' to detect comprehensibility problems in a source questionnaire of a cross-national survey. In: Maksymski, Karin; Gutermuth, Silke; Durand, C. (2005), Measuring interviewer performance in telephone surveys. *Quality and Quantity*, *39*(6), 763–778.

Durand, C. (2005), Measuring interviewer performance in telephone surveys. *Quality and Quantity*, *39*(6), 763–778.

Eckman, S., & Kreuter, F. (2013), Undercoverage rates and undercoverage bias in traditional housing unit listing. *Sociological Methods & Research*, *42*(3), 264–293.

Elliott, M. & West, B. (2015), 'Clustering by interviewer': A source of variance that is unaccounted for in singlestage health surveys. *American Journal of Epidemiology*, 182(2), 118–126.

ESS Sampling Expert Panel (2016), Sampling guidelines: Principles and implementation for the European Social Survey. London: ESS ERIC Headquarters.

Eurofound (2017), *European Quality of Life Survey 2016: Quality of life, quality of public services, and quality of society*, Publications Office of the European Union, Luxembourg.

European Social Survey (2014), Weighting European Social Survey data. London: ESS ERIC Headquarters.

European Social Survey (2015), Round 8 survey specification for ESS ERIC member, observer, and guest countries. London: ESS ERIC Headquarters.

European Social Survey (2016a), ESS Round 8 interviewer briefing: Interviewer manual. London: ESS ERIC Headquarters.

European Social Survey (2016b), ESS Round 8 interviewer briefing: NC Manual. London: ESS ERIC Headquarters.

European Social Survey (2016c), ESS Round 8 translation guidelines. London: ESS ERIC Headquarters.

European Social Survey (2016d), ESS8 - 2014 Documentation Report. Edition 3.1. Retrieved 1 March 2018 http://www.europeansocialsurvey.org/docs/round7/survey/ESS7_data_documentation_report_e03_1.pdf

European Social Survey (2017a), ESS Round 8 (2016/2017) technical report. London: ESS ERIC.

European Social Survey (2017b), ESS8 - 2016 Documentation Report. Edition 1.0. Retrieved 1 March 2018

http://www.europeansocialsurvey.org/docs/round8/survey/ESS8_data_documentation_report_e01_0.pdf

European Union (2015), *ESS handbook for quality reports*, 2014 edition. Luxembourg: Publications Office of the European Union.

Eurostat (2015), Quality assurance framework of the European Statistical System, Version 1.2. Retrieved 1 March 2018, http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646 .

Eurostat (2017), Quality report of the European Union Labour Force Survey, 2015. Luxembourg: Publications Office of the European Union.

Forsman, G. (1993), Sampling individuals within households in telephone surveys. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1113–1118.

Fowler, F. J. Jr. (1995), *Improving survey questions: Design and evaluation* (*Vol. 38, Applied social research methods series*), Thousand Oaks, CA: Sage Publications.

Groves, R. M. (2004), *Survey errors and survey costs* (Vol. 536), John Wiley & Sons.

Groves, R. M. (2006), Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Groves, R. M. (2011), Three eras of survey research. *Public opinion quarterly*, *75*(5), 861-871.

Groves, R. M., & Heeringa, S. G. (2006), Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 439–457.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009), *Survey methodology*. New York, NY: John Wiley & Sons Inc.

Hansen, S.E., Lee, H.J., Lin, Y-c., & McMillan, A. (2016), Instrument technical design. *Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010), Designing questionnaires for multipopulation research. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B-E., Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multicultural, and multiregional contexts* (pp. 33–58), Hoboken, NJ: John Wiley and Sons.

Harkness, J.A., Bilgen, I., Córdova Cazar, A., & Yan, T. (2016), Questionnaire Design. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Heeb, J.-L., & G. Gmel. (2001), ‘Interviewers' and respondents' effects on self-reported alcohol consumption in a Swiss health survey,’ *Journal of Studies on Alcohol* 62, 434–42.

Heeringa, S. G., & O'Muircheartaigh, C. O. (2010), Sampling designs for cross-cultural and cross-national survey programs. In M. Braun, B. Edwards, J. Harkness, T. Johnson, L. Lyberg, P. Mohler, B. E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional and multicultural contexts. 251–267*. Hoboken, NJ: Wiley.

Heeringa, S. G., Wells, J. E., Hubbard, F., Mneimneh, Z. N., Chiu, W. T., Sampson, N. A., & Berglund, P. A. (2008), Sample designs and sampling procedures. In R. Kessler & T.B. Üstün, *The WHO World Mental Health Surveys: Global perspectives on the epidemiology of mental disorders*, 14–32.

Heeringa, S.G., West, B.T., & Berglund, P.A. (2010), Applied survey data analysis. Chapman and Hall, London.

Henry, K. & Valliant, R. (2012), Comparing alternate weighting strategies. Proceedings of the Survey Research Methods Section of the American Statistical Association.

Hill, T., & Westbrook, R. (1997), SWOT analysis: It's time for a product recall. *Long range planning, 30*(1), 46-52.

Hofer, C.W. & Schendel, D. (1978), *Strategy formulation: Analytical concepts*. St. Paul, MN: West Publishing Company.

Hubbard, F., Lin, Y-c., Zahs, D., & Hu, M. (2016), Sample design. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003), Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public opinion quarterly, 67*(1), 79–125.

Hox, J. J., & De Leeuw, E. D. (1994), A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity*, *28*(4), 329–344.

Hyder, S., Bilal, L., Akkad, L., Lin, Y. C., Al-Habeeb, A., Al-Subaie, A., ... & Altwaijri, Y. (2017), Evidence-based guideline implementation of quality assurance and quality control procedures in the Saudi National Mental Health Survey. *International journal of mental health systems*, *11*(1), 60.

International Monetary Fund (2012), Data quality assessment framework. http://dsbb.imf.org/pages/dqrs/DQAF.aspx. Accessed 1 February 2018.

Jäckle, A., Lynn, P., Sinibaldi, J., & Tipping, S. (2013), The effect of interviewer personality, skills and attitudes on respondent co-operation with face-to-face surveys. S*urvey research methods 7*(2), 1-15.

Japec, L. (2005), Interviewer burden and its effects on data quality in the Swedish part of the European social survey. Dissertation, Stockholm University, Faculty of Social Sciences, Department of Statistics.

Jann, B., & Hinz, T. (2016), Research question and designs for survey research. In C. Wolf, D. Joye, T.W. Smith, & Y. Fu (Eds.) *The Sage handbook of survey methodology.*, Los Angeles et al: Sage Publications, 105-121.

Johnson, T. P., & Parsons J.A. (1994), Interviewer effects on self-reported substance use among homeless persons. *Addictive behaviors, 19*(1), 83–93.

Kalton, G. (1983), *Introduction to survey sampling.* Newbury Park, CA: Sage Publications.

Kalton, G., & Flores-Cervantes, I. (2003), Weighting methods, *Journal of official statistics, 19*(2), 81-97.

Kessler, R. C., Ustun, T. B., & World Health Organization (2008), *The WHO world mental health surveys: Global perspectives on the epidemiology of mental disorders.* Cambridge; New York: Geneva: Cambridge University Press; Published in collaboration with the World Health Organization.

Kirgis, N., & Lepkowski, J. (2013), Design and management strategies for paradata- driven responsive design: Illustrations from the 2006-2010 National Survey of Family Growth. In F. Kreuter, (Ed.), *Improving surveys with paradata: Analytic uses of process information*. Hoboken, NJ: John Wiley and Sons.

Kish, L. (1965), *Survey sampling*. New York, NY: John Wiley & Sons

Koch, A. (2016), Assessment of socio-demographic sample composition in ESS Round 6. Mannheim: European Social Survey, GESIS.

Koch, A. (2018), Within-household selection of respondents. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology*. New York, NY: John Wiley & Sons, Inc.

Koch, A., Fitzgerald, R., Halbherr, V. & Villar, A. (2016), ESS Round 8 guidelines on fieldwork progress reporting. London: ESS ERIC Headquarters

Kolarz, P., Angelis, J., Krčál, A., Simmonds, P., Traag, V., & Wain, M. (2017), Comparative impact study of the European Social Survey (ESS) ERIC. Technopolis Group.

Kolczynska, M., & Schoene, M. (2018), Survey data harmonization and the quality of data documentation in cross-national surveys. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology*. New York, NY: John Wiley & Sons Inc.

Kotler, P. (2000), *Marketing management.* Prentice-Hall, Upper Saddle River, NJ.

Kreuter, F. (2013a), *Improving surveys with paradata*. Hoboken, NJ: John Wiley & Sons, Inc.

Kreuter, F. (2013b), Facing the nonresponse challenge. *Annals of the American academy of political and social sciences, 645*, 23–35.

Kreuter, F., & Olson, K. (2013), *Paradata for nonresponse error investigation*. University of Nebraska - Lincoln, Sociology Department, Faculty Publications. Paper 220.

Laflamme, F., & St-Jean, H. (2011), Proposed indicators to assess interviewer performance in CATI survey. *Proceedings of the Joint Statistical Meetings*, 118–28.

Lavrakas, P. J. (2008), Within-household respondent selection: How best to reduce total survey error? MRC Respondent Selection Report.

Lipps, O. (2007), Interviewer and respondent survey quality effects in a CATI Panel. Bulletin of Sociological Methodology, 95, 5–25.

Lohr, S. (2009), *Sampling: design and analysis*. Nelson Education.

Loosveldt, G. & Beullens, K. (2014), Report on interviewer-related variances in the European Social Survey Round 6 (ESS ERIC Deliverable 7.2), KU Leuven: Centre for Sociological Research.

Lyberg, L. (2012), Survey quality. *Survey Methodology, 38*(2), 107–130.

Lyberg, L. (2016), Prevailing issues and the future of comparative surveys. Paper presented at the 2nd International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts, July 26-29, Chicago, U.S.A.

Lyberg, L. E., & Biemer, P. P. (2008), Quality assurance and quality control in surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology*. New York/London: Lawrence Erlbaum Associates.

Lyberg, L. E., & Stukel, D. M. (2010), Quality assurance and quality control in cross-national comparative studies. In J. A. Harkness, M. Braun, B. Edwards, T. Johnson, L. E. Lyberg, P. Ph. Mohler,

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

B-E., Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multicultural and multiregional contexts* (pp. 227-249), Hoboken, NJ: John Wiley & Sons.

Lynn, P., Häder, S., Gabler, S., & Laaksonen, S. (2007), Methods for achieving equivalence of samples in cross-national surveys: The European Social Survey experience. *Journal of Official Statistics*, *23*, 107–124.

Maitland, A., & Presser, S. (2017), How do question evaluation methods compare in predicting problems observed in typical survey conditions? *Survey Statistics and Methodology. 0*, 1–26.

Marker, D. A., and Morganstein, D. R. (2004), Keys to successful implementation of continuous quality improvement in a statistical agency. *Journal of Official Statistics, 20*(1), 125–136.

McDonald, M. (1999), Marketing plans, Butterworth-Heinemann Press, Oxford.

Meitinger, K. & Behr, D. (2016), Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods*, *28*(4), 363–380.

Miller, K. (2018), Conducting cognitive interviewing studies to examine survey question comparability. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology*. New York, NY: John Wiley & Sons Inc.

Mitchell, R. J., Richardson, E. A., Shortt, N. K., & Pearce, J. R. (2015), Neighborhood environments and socioeconomic inequalities in mental well-being. *American journal of preventive medicine*, *49*(1), 80–84.

Mneimneh, Z., de Jong, J., & Moaddel, M. (2017), Toward a better understanding of the effect of interviewers' attitudes on reporting sensitive religious information. Paper presented at the European Social Research Association, Lisbon, July 18 – 21.

Mneimneh, Z., Lyberg, L., Sharma, S., Vyas, M., Sathe, D.B., Malter, F., & Altwain. (2018), Case Studies on Monitoring Interviewer Behavior in International and Multinational Surveys. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology*. New York, NY: John Wiley & Sons Inc.

Mohler, P. Ph. (2006), Sampling from a universe of items and the de-machiavellization of questionnaire design. In M. Braun & P. Ph. Mohler (Eds.), *Beyond the horizon of measurement - Festschrift in honor of Ingwer Borg* (ZUMA-Nachrichten Spezial,10), Mannheim, Germany: ZUMA.

Murphy, J., M. Keating, and Edgar, J. (2013), Crowdsourcing in the cognitive interviewing process. Paper presented at the FCSM Research Conference, Washington, DC, November 4–6.

O'Muircheartaigh, C., & Campanelli, P. (1998), The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *161*(1), 63–77.

Pennell, B.-E., Cibelli Hibben, K. L., Lyberg, L., Mohler, P. Ph., and Worku, G. (2017), A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg … B. West (Eds.), *Total survey error in practice*. New York: John Wiley and Sons.

Petrakos, M., Kleideri, M., & Ieromnimon, A. (2011), Quality assessment of the 2nd European Quality of Life Survey. Report published by European Foundation for the Improvement of Living and Working Conditions.

Peytchev, A. (2013), Consequences of survey nonresponse. *The ANNALS of the American academy of political and social science*, 645(1), 88–111.

Peytcheva, E. & Groves, R. M. (2009), Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193–201.

PIAAC (2010), PIAAC Reducing Nonresponse Bias and Preliminary Nonresponse Bias Analysis. OMB# 1850-0870 v.5.

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

PIAAC (2014), Technical standards and guidelines. Retrieved from http://www.oecd.org/skills/piaac/PIAAC-NPM(2014_06)PIAAC_Technical_Standards_and_Guidelines.pdf on 1 February 2018.

Pickery, J., & Loosveldt, G. (2000), Modeling interviewer effects in panel surveys:-an application. *Survey Methodology*, *26*(2), 189-198.

Pickery, J., & Loosveldt, G. (2002), A multilevel multinomial analysis of interviewer effects on various components of unit nonresponse. *Quality & Quantity, 36*, 427–437.

Pierchala, C. E., & Surti, J. (2009), Control charts as a tool for data quality control. *Journal of Official Statistics*, 25(2), 167–191.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. Ò., Martin, E. A., Martin, J., et al. (Eds.), (2004), *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: John Wiley and Sons.

Rizzo, L., Brick, J. M., & Park, I. (2004), A minimally intrusive method for sampling persons in random digit dial surveys. *Public Opinion Quarterly, 68*, 2, 267–274.

Robbins, M. (2018), New frontiers in detecting data fabrication. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), Advances in comparative survey methodology. New York, NY: John Wiley & Sons Inc.

Saris, W. E., & Gallhofer, I. N. (2007), *Design, evaluation, and analysis of questionnaires for survey research* (1st edition), John Wiley & Sons.

Saris, W. E., & Gallhofer, I. N. (2014), *Design, evaluation, and analysis of questionnaires for survey research* (2nd edition), John Wiley & Sons.

Scherpenzeel, A., Maineri, A., Bristle, J., Pflüger, S.M., Mindarova, I., Butt, S., Zins, S., Emery, T. & Luijkx, R., (2017), Report on the use of sampling frames in European studies. *SERISS-Deliverable*, (2.1),

Schnaars, S.P. (1998), *Marketing strategy*. Free Press, New York, NY.

Shackman, G. (2001), Sample size and design effect. Technical report, Albany Chapter of American Statistical Association.

Singer, E. (2002), The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (163–178), New York, NY: John Wiley & Sons.

Sinibaldi, J., Durrant, G. B., & Kreuter, F. (2013), Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly*, *77*(S1), 173–193.

Smith, T. W. (2004), Developing and evaluating cross-national survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 431–452), Hoboken, New Jersey: John Wiley & Sons.

Smith, T. W. (2007), Survey nonresponse procedures in cross-national perspective: The 2005 ISSP nonresponse survey. *Survey Research Methods, 1*(1), 45–54.

Smith, T. W. (2011), Refining the total survey error perspective. *International Journal of Public Opinion Research*, *23*(4), 464–484.

Smith, T.W. (2018), Improving Multinational, Multiregional and Multicultural Comparability (3MC) Using the Total Survey Error (TSE) Paradigm. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology*. New York, NY: John Wiley & Sons Inc.

Statistisches Bundesamt (Ed.) (2012), Geburten in Deutschland. Wiesbaden: Ausgabe 2012.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010a), *Improving survey response. Lessons learned from the European Social Survey*. Chichester: John Wiley and Sons, Ltd.

Stoop, I., Matsuo, H., Koch, A., & Billiet, J. (2010b), Paradata in the European Social Survey: Studying nonresponse and adjusting for bias. *Proceedings of the Survey Research Methods section, ASA,* 407–421.

Stoop, I., Koch, A., Halbherr, V., Loosveldt, G., & Fitzgerald, R. (2016), Field procedures in the European Social Survey Round 8: Guidelines for enhancing response rates and minimising nonresponse bias. London: ESS ERIC Headquarters.

Sudman, S. & Bradburn, N. M. (1982), *Asking questions: A practical guide to questionnaire design.* San Francisco, CA: Jossey-Bass.

Survey Research Center. (2016), *Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved 18 April 2017, from http://www.ccsg.isr.umich.edu/.

Thompson, K. J., & Oliver, B. E. (2012), Response rates in business surveys: Going beyond the usual performance measure. Journal of Official Statistics, 28(2), 221–237.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000), *The psychology of survey response*. Cambridge University Press.

Ustun, T. B., Chatterji, S., Mechbal, A., & Murray, C. J. L. (2005), Chapter X: Quality assurance in surveys: Standards, guidelines, and procedures. In United Nations Statistical Division, United Nations Department of Economic and Social Affairs (Eds.), *Household surveys in developing and transition countries*. New York, NY: United Nations.

Valliant R., Dever J.A., & Kreuter F. (2013), Calibration and other uses of auxiliary data in weighting. in: practical tools for designing and weighting survey samples. *Statistics for social and behavioral sciences*, 349–395. Springer, New York, NY.

van den Brakel, J. A., Vis-Visschers, R., & Schmeets, J. J. G. (2006), An experiment with data collection modes and incentives in the Dutch Family and Fertility Survey for Young Moroccans and Turks. *Field methods, 18,* 321–334.

Van Oostrum, T., Sandor, E., & van Houten, G. (2017), Effects of fieldwork delay on estimates of subjective wellbeing in the 4th European Quality of Life Survey. Paper presented at the Comparative Survey Design and Implementation Workshop, Mannheim, March 16-18.

Vassallo, R., Durrant, G. B., Smith, P. W. F., & Goldstein, H. (2015), Interviewer effects on nonresponse propensity in longitudinal surveys: a multilevel modelling approach. *Journal of the Royal Statistical Society Series A, 178*(1), 83–99.

Vehovar, V., Slavec, A., & Berzelak, N. (2012), Costs and errors in fixed and mobile phone surveys. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 277–295),

Vila, J., Cervera, J.L., & Carausu, F. (2013), Quality assessment of the third European Quality of Life Survey. Report published by European Foundation for the Improvement of Living and Working Conditions.

Vila, J., & Cervera, J.L. (2014), Revision of the weighting strategy in the European Quality of Life Survey (EQLS), Report published by European Foundation for the Improvement of Living and Working Conditions.

Villar, A., & Fitzgerald, R. (2017), Using mixed modes in survey research. *Values and identities in Europe: Evidence from the European Social Survey, 202*–273.

Villar, A., Sommer, E., Finnøy, D., Johannesen, B.-O., Soidla, I., Humphrey, A., Kurdija, S., Ainsaar, M., Vovk, T., & Berzelak, N. (2017), Design and recruitment of a probability based CROss-National Online Survey (CRONOS) panel. Paper presented at the Comparative Survey Design and Implementation Workshop, Mannheim, March 16-18.

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012), Use of paradata in a responsive design framework to manage a field data collection. *Journal of official statistics, 28*(4), 477–499.

Wagner, J., & Stoop, I. (2018), Comparing nonresponse and nonresponse biases in multinational, multiregional and multicultural contexts. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology*. New York, NY: John Wiley & Sons, Inc.

West, B. T., & Groves, R. M. (2013), A propensity-adjusted interviewer performance indicator. *Public opinion quarterly, 77*(1), 352–374

West, B. T., & Kreuter, F. (2013), Factors affecting the accuracy of interviewer observations: Evidence from the National Survey of Family Growth. *Public opinion quarterly*, *77*(2), 522–548.

Yasukawa, K., Hamilton, M., & Evans, J. (2017), A comparative analysis of national media responses to the OECD Survey of Adult Skills: policy making from the global to the local? *Compare: A journal of comparative and international education*, *47*(2), 271–285.

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

Annexes

Annex 1: Compliance with additional QAP indicators

Table A1. Additional Sampling Frame Development Quality Indicators

Coherence & Comparability			
	Indicator	Target	UM Assessment
1	RW: Percentage of countries where the specified information on degree of urbanisation uses a common set of categories	100%	Target not met
2	RW: Percentage of countries where a common set of variables are used for stratification	100%	Target not met
Accessibility			
3	RQ: Enumeration plan is provided (in countries where enumeration occurs)	100%	Target met
4	RQ: Distributions across stratification categories of reference statistics and selected PSUs/respondents are provided	100%	Target met
5	RQ: Characteristics of the sampling frame and procedure are documented in complete accordance with the template	100%	Target met
6	RQ: Characteristics of the reference statistics are documented in complete accordance with the template	100%	Target met
7	RQ: All stratification variables and distributions of universe statistics are made available in interim and final datasets	100%	Target met[47]
Punctuality			
8	RQ: Enumeration finalised before fieldwork	Yes	Target met
9	RQ: Quality check on enumeration finalised before fieldwork	Yes	Target met
10	RQ: Sampling plan delivered at agreed date from KP to EF	Yes	Target not met
11	RQ: Sampling plans approved at agreed date by EF	Yes	Target not met
12	RQ: Gross sample provided to national agencies at agreed date	Yes	Target not met
13	RQ: Training of enumerators delivered at agreed date	Yes	N/A

[47] Variables were not included in the interim dataset as weights were not yet complete. However, all such variables were included in the final dataset and therefore we consider there to be compliance with regards to data quality.

Table A2. Additional Questionnaire Development Quality Indicators

Relevance and Timeliness			
	Indicator	Target	UM Assessment
14	RQ: Questionnaire consulted by Eurofound's stakeholders/Advisory Committee	Yes	Target met
Accessibility			
15	RQ: Comprehensive documentation of the process of advance translation	Yes	Target not met
16	RQ: Clear translation instructions for advance translation	Yes	Target not met
17	RQ: Percentage of questionnaire items for which systematic documentation is provided about the extent to which answers in the cognitive interviews correspond with intended concepts	100%	Target not met
18	RQ: Translation materials are constructed using input from the cognitive test and advance translation, are provided to the translators, and are made publicly available.	Yes	Target met
19	RQ: Percentage of countries for which systematic documentation of results of initial translation (in accordance with template) is provided	100%	Target met
20	RQ: Percentage of countries for which systematic documentation in English is provided about the process and results of adjudication (in accordance with template)	100%	Target met
21	RQ: Percentage of countries for which systematic documentation in English is provided about the process and results of the cross-national review (per template)	100%	Target met
Punctuality			
22	RQ: Timeline for questionnaire development is defined and kept	Yes	Target met
23	RQ: Advance translation delivered at agreed date	Yes	Target met
24	RQ: Cognitive test delivered at agreed date	Yes	Target met
25	RQ: Selecting of questions eligible for translation delivered at agreed date	Yes	Target met
26	RQ: Initial translation delivered at agreed date	Yes	Target met
27	RQ: Within country adjudication (overall) delivered at agreed date	Yes	Target met
28	RQ: Cross country review (overall) delivered at agreed date	Yes	Target not met
29	RQ: Final translated questionnaires (language version) delivered at agreed date	Yes	Target not met

Table A3. Additional Fieldwork Implementation Quality Indicators

Accessibility

	Indicator	Target	UM Assessment
30	RQ: All advance letters and promocards are made available from Eurofound website	Yes	Target met
31	RQ: Percentage of countries for which all training materials are provided	100%	Target met
32	RQ: Percentage of countries covered in weekly monitoring data (per the template)	100%	Target not met
33	RQ: Comprehensive methodological and fieldwork report provided	Yes	Target met
Punctuality			
34	RQ: Meeting of national fieldwork managers held before start of fieldwork	Yes	Target met
35	RQ: Interviewer training delivered before start of fieldwork	Yes	Target met
36	RQ: Interviewer training materials delivered at agreed date	Yes	Target not met
37	RQ: CAPI/data entry process programmed and finalised at agreed date	Yes	Target not met
38	RQ: Number of times that the weekly monitoring data for the preceding week is not delivered on Tuesday by EOB	0	Target not met
39	RQ: Number of times that the quantitative indicators in the weekly monitoring data and the progress and projections not checked by EF the following Thursday EOB	0	Target not met
40	RQ: Number of days that fieldwork continues after the agreed date	0	Target not met
41	RQ: Technical and fieldwork report delivered at agreed date	Yes	Target not met

Table A4. Additional Weighting Quality Indicators

Coherence and Comparability			
	Indicator	Target	UM Assessment
42	RQ: Weighting strategy includes references to academic literature demonstrating that the selection of weighting variables and procedures takes common practice of weighting in international surveys into account	Yes	Target met
Accessibility			
43	RQ: Percentage of countries for which the weighting strategy and procedure are made completely transparent in the weighting report	100%	Target met
44	RQ: Design weight included in dataset	Yes	Target met
45	RQ: Procedure for constructing design weights outlined in weighting report	Yes	Target met
46	RQ: Post-stratification weight included in dataset	Yes	Target met
47	RQ: Procedure for constructing post-stratification weights in weighting report	Yes	Target met

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

48	RQ: Supra-national weights included in dataset	Yes	Target met
49	RQ: Procedure for construction & sources for supra-national weights described in weighting report	Yes	Target met
50	RQ: Trimmed and untrimmed weights are included in the dataset	Yes	Target met
51	RQ: Trimming cut-off points and number of trimmed cases for each country are included in the weighting report	Yes	Target met
Punctuality			
52	RQ: Weighting strategy delivered at agreed date	Yes	Target not met
53	RQ: Design weights delivered at agreed date	Yes	Target not met
54	RQ: Post-stratification weights delivered at agreed date	Yes	Target not met
55	RQ: Supra-national weights delivered at agreed date	Yes	Target not met

Annex 2. Questionnaire criteria

Non-sensitive questions about behaviour

- With closed questions, include all reasonable possibilities as explicit response options[a]
- Make questions as specific as possible[a]
- Use words that virtually all respondents will understand[a]
- Lengthen the questions by adding memory cues to improve recall[a]
- When forgetting is likely, use aided recall[a]
- When the events of interest are frequent but not very involving, have respondents use a diary[a]
- When long recall periods must be used, use a life event calendar to improve reporting[a]
- To reduce telescoping errors, ask respondents to use household records or use bounded recall (or do both) [a]
- If cost is a factor, consider whether proxies might be able to provide accurate information[a]
- Use a WH-interrogative (what, when, etc.) in a direct request to avoid leading or unbalanced requests[b]
- When asking about frequency of events, better estimation occurs with a two-part question, with the first part asking a about a long time period (e.g., 5 years), and the second asking about the shorter time period (e.g., 1 year)[b]
- Use two-part questions, asking first about whether something occurs, and then second about frequency, to avoid that an item has a hidden assumption about occurence[b]

Sensitive questions about behaviour

- Use open rather than closed questions for eliciting the frequency of sensitive behaviours[a]
- Use long rather than short questions[a]
- Use familiar words in describing sensitive behaviours[a]
- Deliberately load the question to reduce misreporting[a]
- Ask about long periods (such as one’s entire lifetime) or periods from the distant past first in asking about sensitive behaviours[a]
- Embed the sensitive question among other sensitive items to make it stand out less[a]
- Use self-administration or some similar method to improve reporting[a]
- Consider collecting the data in a diary[a]
- At the end of the questionnaire, include some items to assess how sensitive the key behavioural questions were[a]
- Collect validation data[a]

Attitudinal questions

- Specify the attitude object clearly[a]
- Avoid double-barreled questions[a]
- Measure the strength of the attitude, if necessary using separate items for this purpose[a]
- Use bipolar items except when they might miss key information[a]

- The alternatives mentioned in the question have a big impact on the answers; carefully consider which alternatives to include[a]
- In measuring change over time, ask the same questions each time[a]
- When asking general and specific questions about a topic, ask the general question first[a]
- When asking questions about multiple items, start with the least popular one[a]
- Use closed questions for measuring attitudes[a]
- Use five-point to seven-point response scales and label every scale point[a]
- Ten-point scales should have end-points labeled as fixed reference points[b]
- Start with the end of the scale that is the least popular[a]
- Use analog devices (such as thermometers) to collect more detailed scale information[a]
- Use ranking only if the respondents can see all the alternatives; otherwise, use paired comparisons[a]
- Get ratings for every item of interest; do not use check-all-that-apply items[a]
- Avoid using complex assertions, which refer to cognitive judgment, to precede attitude questions (e.g., ‘Do you think’)[b]
- Avoid conditional clauses [b]
- Avoid questions with hidden assumptions[b]

[a] Denotes questionnaire design criteria as specified in Chapter 7, Groves et al. (2009),

[b] Denotes questionnaire design criteria drawn from principles as discussed in Chapters 3 to 6, Saris and Gallhofer (2014),

Annex 3. Response rates

AAPOR Formulas[48]

Response Rate 1

$RR1 = I/((I+P) + (RNC+O)+(UH+UO))$

Refusal Rate 1

$REF1 = R/((I+P)+(R+NC+O)+(UH+UO))$

Contact Rate 1

$CON1 = (I+P)+R+O/(I+P)+R+O+NC+(UH+UO)$

Cooperation Rate 1

$COOP1 = I/((I+P) + R+O))$

[48] See AAPOR (2016) for formulas.

Table A5. AAPOR Categories and Codes and Response Codes for the 4th EQLS, 3rd EQLS and the ESS

Category		**AAPOR Code**	**Response Codes**		
			4th EQLS	**3rd EQLS**	**ESS**
		1.1	18 Successful interview	Completed interview	V. Number of valid interviews
Interview	**I**	1.2	104 Dropped out (data saved, do not come back)	Partial interview	
		2.11	304 Upfront refusal before household selection	Upfront refusal	
		2.11	305 No one at home (non-final code, appointment possible)	Fixed appointment (no interview)	
		2.11	312 Upfront refusal by another household		
		2.111	303 Refusal by phone (only if confirmed by supervisor)		
		2.111	317 Upfront refusal by another person from selected household		C. Refusal by proxy, or household or address refusal
		2.111	313 Refusal by selected household		
		2.111			D. Refusals by opt-out list
	R	2.112	307 Refusal by selected	Refusal by selected respondent	B. Refusal by respondent
		2.2		Selected respondent currently not at home	
		2.23	320 Unable to enter the building		
		2.24	306 No contact after 4 visits (final code)	No contact	E. No contact (after at least 4 visits)
Eligible, non-interview	**NC**	2.25	308 Away for FW period	Selected respondent away for fieldwork period	M. Respondent emigrated/left the country long term (for more than 6 months)

		2.25	319 Selected respondent moved away		
		2.3			H. Contact, but no interview for other reasons
		2.31	318 Selected respondent deceased		N. Respondent deceased
		2.32		Selected respondent ill at home/hospital	
		2.331	311 Upfront language barrier	Selected respondent has language difficulties	F. Language barrier
		2.332	315 Selected person is physically or mentally unable - incompetent	Selected respondent physically or mentally unable	G. Respondent ill or incapacitated, unable to cooperate throughout fieldwork period
	O	2.332	316 Selected person doesn't speak national languages	Other language	
		3.17	321 Inaccessible/ dangerous	Area inaccessible/dangerous	
	UH	3.18	301 Address Not Found/ Demolished	Address not found/demolished	I. Address not traceable
		3.9	20 System error		U. Invalid interviews
Unknown eligibility, non-interview	**UO**	3.9			Y. Number of sample units not accounted for
		4.0		Ineligible	
		4.5	302 Non-residential address	Non-residential address	J. Address not residential
		4.6	310 Vacant/ empty housing	Vacant property	K. Address not occupied
		4.63			L. Other ineligible address
Not eligible		4.7	322 No eligible respondents		

Table A6. 4th and 3rd EQLS Response, Refusal, Contact, and Cooperation Rates

	4th EQLS				**3rd EQLS**			
Country	**RR1**	**REF1**	**CON1**	**COOP1**	**RR1**	**REF1**	**CON1**	**COOP1**
Albania	0.60	0.29	0.94	0.64	.	.	.	.
Austria	0.34	0.45	0.87	0.39	0.50	0.24	0.77	0.65
Belgium	0.38	0.36	0.82	0.46	0.48	0.36	0.88	0.55
Bulgaria	0.58	0.34	0.95	0.61	0.61	0.19	0.81	0.75
Croatia	0.51	0.31	0.85	0.59	0.46	0.34	0.81	0.57
Cyprus	0.52	0.29	0.89	0.58	0.78	0.16	0.97	0.81
Czech Rep	0.59	0.25	0.87	0.68	0.45	0.40	0.86	0.52
Denmark	0.35	0.45	0.86	0.40	0.30	0.45	0.76	0.39
Estonia	0.43	0.31	0.77	0.55	0.54	0.19	0.74	0.73
Finland	0.34	0.39	0.76	0.45	0.39	0.36	0.77	0.50
FRYOM Macedonia	0.64	0.27	0.96	0.66	0.78	0.21	0.99	0.79
France	0.30	0.33	0.72	0.42	0.30	0.36	0.68	0.45
Germany	0.18	0.67	0.96	0.19	0.41	0.47	0.89	0.46
Greece	0.25	0.52	0.83	0.30	0.43	0.52	0.98	0.44
Hungary	0.55	0.36	0.94	0.58	0.42	0.55	0.97	0.43
Ireland	0.48	0.26	0.80	0.60	0.53	0.16	0.72	0.74
Italy	0.26	0.59	0.89	0.30	0.40	0.49	0.90	0.44
Kosovo	.	.	.	.	0.90	0.10	1.00	0.90
Latvia	0.45	0.33	0.80	0.56	0.46	0.18	0.64	0.71

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

Lithuania	0.38	0.27	0.70	0.55	0.45	0.29	0.75	0.60
Luxembourg	0.22	0.45	0.74	0.30	0.15	0.53	0.72	0.20
Malta	0.51	0.31	0.84	0.61	0.69	0.19	0.88	0.79
Montenegro	0.71	0.15	0.89	0.80	0.45	0.54	0.99	0.45
Netherlands	0.29	0.47	0.82	0.35	0.31	0.45	0.79	0.39
Poland	0.33	0.53	0.87	0.38	0.61	0.27	0.88	0.70
Portugal	0.52	0.36	0.91	0.57	0.38	0.23	0.62	0.61
Romania	0.55	0.23	0.82	0.67	0.59	0.17	0.77	0.77
Serbia	0.69	0.24	0.95	0.72	0.46	0.46	0.92	0.50
Slovakia	0.56	0.25	0.88	0.63	0.62	0.24	0.86	0.72
Slovenia	0.44	0.34	0.85	0.52	0.48	0.37	0.86	0.56
Spain	0.50	0.22	0.75	0.67	0.34	0.41	0.76	0.45
Sweden	0.16	0.20	0.64	0.24	0.46	0.44	0.90	0.51
Turkey	0.68	0.14	0.87	0.79	0.53	0.33	0.85	0.62
UK	0.31	0.46	0.80	0.38	0.26	0.38	0.66	0.39

Table A7. ESS Round 7 and 8 Response, Refusal, Contact, and Cooperation Rates

	Round 8				Round 7			
Country	**RR1**	**REF1**	**CON1**	**COOP1**	**RR1**	**REF1**	**CON1**	**COOP1**
Austria	0.53	0.38	0.92	0.57	0.51	0.33	0.87	0.60
Belgium	0.54	0.23	0.88	0.61	0.53	0.25	0.92	0.58
Czech Rep	0.69	0.28	0.98	0.70	0.69	0.27	0.97	0.71

Denmark	.	.	.	.	0.49	0.33	0.91	0.54
Estonia	0.66	0.16	0.86	0.77	0.57	0.19	0.83	0.69
Finland	0.55	0.25	0.89	0.62	0.60	0.23	0.91	0.66
France	0.52	0.21	0.80	0.64	0.49	0.23	0.79	0.61
Germany	0.30	0.48	0.86	0.34	0.30	0.46	0.89	0.34
Hungary	.	.	.	.	0.51	0.31	0.88	0.58
Iceland	0.44	0.37	0.88	0.50	.	.	.	.
Ireland	0.61	0.16	0.86	0.71	0.58	0.22	0.90	0.64
Israel	0.74	0.19	0.95	0.77	0.74	0.11	0.88	0.84
Lithuania	.	.	.	.	0.69	0.24	0.94	0.73
Netherlands	0.52	0.32	0.91	0.57	0.60	0.29	0.94	0.64
Norway	0.50	0.24	0.88	0.57	0.52	0.27	0.89	0.58
Poland	0.62	0.14	0.81	0.77	0.58	0.17	0.81	0.72
Portugal	.	.	.	.	0.46	0.37	0.96	0.48
Russian Federation	0.63	0.24	0.90	0.70	.	.	.	.
Slovenia	0.54	0.34	0.92	0.59	0.51	0.33	0.90	0.57
Spain	.	.	.	.	0.64	0.13	0.82	0.78
Sweden	0.41	0.38	0.92	0.45	0.47	0.34	0.93	0.51
Switzerland	0.50	0.25	0.87	0.58	0.51	0.25	0.87	0.58
UK	0.43	0.34	0.85	0.50	0.43	0.33	0.84	0.51

Annex 4. Comparisons of sample composition data[49]

Table A8: Differences in the proportional distribution of gender/age between 4th EQLS and Eurostat

	Male						Female					
	18-29	30-39	40-49	50-59	60-69	70+	18-29	30-39	40-49	50-59	60-69	70+
Austria	-5.1%	-0.2%	-0.2%	-1.3%	-0.3%	-3.8%	-0.5%	8.5%	4.8%	2.0%	1.7%	-5.6%
Belgium	-1.9%	0.8%	-1.5%	0.9%	1.0%	1.7%	-1.5%	-1.2%	-1.1%	0.7%	1.1%	1.1%
Czech Republic	-4.1%	-2.2%	-1.6%	1.3%	2.8%	-0.5%	-1.5%	0.8%	1.6%	0.5%	4.8%	-1.9%
Estonia	-5.3%	-3.7%	-2.6%	-0.2%	0.2%	1.5%	-3.8%	-1.5%	0.6%	2.9%	6.0%	5.8%
Finland	-5.5%	-3.3%	-2.2%	1.0%	5.6%	5.8%	-4.2%	-3.0%	-2.2%	0.3%	4.3%	3.3%
Germany	-3.0%	-0.9%	-2.2%	0.1%	2.8%	-2.0%	-1.3%	2.1%	2.2%	3.4%	2.2%	-3.4%
Ireland	-3.7%	-3.7%	-3.2%	-1.0%	3.5%	4.0%	-3.6%	1.7%	2.5%	0.8%	0.9%	1.9%
Latvia	-5.1%	-2.8%	-2.7%	-1.8%	0.9%	2.2%	-2.4%	-2.0%	-0.4%	2.6%	3.6%	7.9%
Lithuania	-3.0%	-2.0%	-3.5%	-1.3%	1.0%	2.5%	-1.3%	-0.4%	-1.4%	1.2%	2.4%	5.7%
Malta	-4.9%	-2.0%	-2.1%	-0.1%	0.0%	3.3%	-3.8%	0.1%	0.5%	1.2%	2.4%	5.4%
Netherlands	-3.4%	-1.3%	-2.3%	-0.1%	1.3%	0.7%	-4.2%	0.4%	1.6%	1.5%	2.0%	3.8%
Poland	-5.4%	-5.5%	-1.6%	-1.5%	2.7%	1.7%	0.2%	-1.1%	1.2%	5.7%	4.5%	-0.9%
Slovakia	-6.3%	-3.4%	-2.2%	-0.4%	3.9%	1.0%	-5.6%	-2.4%	-0.2%	4.3%	6.6%	4.7%
Slovenia	-1.6%	-3.6%	-1.3%	0.8%	2.2%	1.0%	-2.3%	-1.5%	0.3%	1.6%	2.9%	1.6%
Spain	-0.2%	-2.4%	0.4%	0.6%	1.3%	-1.7%	-1.0%	-0.1%	1.8%	1.9%	2.2%	-2.8%
Sweden	-4.5%	-1.3%	-2.2%	-0.3%	3.7%	5.7%	-5.2%	-1.9%	-1.9%	-1.0%	3.8%	5.1%

[49] In each table in Annex 4, countries are sorted alphabetically as follows: 1) EU register countries; 2) EU enumeration countries; and 3) EU candidate countries, with Turkey, as the lone register country, listed first among them.

UK	-4.2%	-0.7%	-1.2%	-1.1%	1.5%	2.7%	-1.8%	2.1%	-0.9%	-0.8%	1.5%	2.9%
Bulgaria	-4.4%	-2.4%	-0.9%	0.9%	1.2%	0.2%	-1.2%	-0.8%	0.3%	0.5%	4.8%	1.8%
Croatia	-3.4%	-1.2%	-0.9%	-2.5%	-0.4%	-2.0%	-0.4%	5.0%	2.7%	2.6%	0.9%	-0.3%
Cyprus	-7.1%	-2.4%	-2.3%	-2.5%	2.0%	5.0%	-6.2%	0.1%	0.1%	3.3%	3.2%	6.8%
Denmark	-3.3%	-0.9%	0.2%	0.8%	0.5%	3.1%	-3.3%	-0.6%	-0.9%	0.5%	0.5%	3.3%
France	-3.3%	1.1%	0.4%	-1.4%	1.1%	-0.4%	-1.4%	1.1%	3.3%	0.1%	1.4%	-2.0%
Greece	-2.7%	-2.3%	-0.3%	-0.8%	-0.3%	0.6%	-1.7%	1.2%	2.4%	2.7%	2.1%	-0.9%
Hungary	-5.5%	-3.0%	-2.0%	-0.6%	0.9%	2.2%	-2.5%	-0.8%	0.8%	1.8%	4.2%	4.4%
Italy	-3.9%	-2.3%	-0.8%	-0.6%	0.7%	-0.5%	-2.2%	3.0%	3.2%	3.3%	2.8%	-2.6%
Luxembourg	-3.1%	-1.1%	-2.0%	-1.1%	1.4%	0.4%	-2.6%	2.8%	3.9%	0.8%	1.4%	-0.7%
Portugal	-2.5%	-3.1%	-1.0%	0.9%	0.6%	1.9%	-1.6%	-1.1%	-0.2%	0.5%	3.4%	2.2%
Romania	-4.2%	-4.1%	-1.0%	-1.5%	1.4%	0.0%	-2.3%	-1.7%	2.7%	1.9%	4.7%	4.1%
Turkey	4.0%	2.4%	-1.6%	-2.2%	-0.5%	-1.6%	4.4%	2.3%	-1.2%	-1.4%	-2.0%	-2.7%
Albania	-4.1%	-3.4%	-2.1%	-1.9%	2.1%	2.0%	-1.8%	0.3%	1.1%	3.3%	4.3%	0.2%
FYR Macedonia	-3.7%	-3.2%	-0.3%	-0.3%	4.2%	1.0%	-3.4%	-0.9%	-0.1%	-0.1%	4.2%	2.7%
Montenegro	4.5%	-1.6%	-1.0%	2.3%	-1.2%	-2.7%	4.8%	-0.9%	3.0%	-0.7%	-3.0%	-3.6%
Serbia	5.9%	9.0%	9.0%	10.1%	8.1%	4.6%	7.4%	11.1%	10.3%	9.9%	8.5%	6.1%

Table A9: Differences in the unweighted proportional distribution of gender/age between 4th EQLS and ESS Round 7

	Male						Female					
	18-29	30-39	40-49	50-59	60-69	70+	18-29	30-39	40-49	50-59	60-69	70+
Austria	-3.0%	0.0%	0.3%	-1.5%	-1.6%	-4.4%	0.6%	7.4%	4.3%	1.3%	1.7%	-5.2%
Belgium	-2.2%	1.3%	-1.0%	0.4%	-0.5%	1.7%	-1.9%	-0.4%	-1.8%	0.7%	0.1%	3.6%
Czech Republic	-2.6%	-0.9%	-1.8%	1.1%	2.6%	0.0%	-1.0%	-0.6%	-0.5%	-1.4%	3.9%	1.2%
Estonia	-3.8%	-1.7%	-1.0%	0.6%	0.2%	1.1%	-1.9%	-2.5%	0.3%	1.2%	4.3%	3.2%
Finland	-3.9%	-1.3%	-1.6%	0.2%	2.9%	4.6%	-1.8%	-2.3%	-1.9%	-0.4%	2.1%	3.4%
Germany	-1.8%	-0.3%	-2.0%	-1.5%	0.5%	-2.3%	-0.4%	2.4%	2.7%	2.7%	1.3%	-1.3%
Ireland	-1.9%	-0.9%	-0.3%	-1.0%	1.0%	1.6%	-1.6%	2.2%	2.5%	0.3%	-2.2%	0.2%
Lithuania	1.4%	0.3%	-2.7%	-1.2%	0.6%	2.3%	0.2%	-1.2%	-2.5%	-1.0%	-1.3%	5.0%
Netherlands	-0.2%	0.8%	0.3%	0.8%	-0.8%	-1.6%	-2.2%	-0.9%	0.8%	-0.7%	0.0%	3.7%
Poland	-5.4%	-3.5%	-1.3%	-0.4%	2.0%	1.0%	0.5%	-0.8%	0.7%	5.5%	2.8%	-1.2%
Slovenia	-1.2%	-0.7%	1.9%	0.9%	0.3%	-0.4%	-2.9%	-0.8%	0.2%	0.2%	1.1%	1.6%
Spain	-0.7%	-2.2%	0.1%	-0.1%	0.6%	-2.6%	-1.0%	0.9%	2.8%	1.6%	2.4%	-1.7%
Sweden	-3.5%	-0.2%	-1.7%	0.4%	2.6%	3.6%	-4.0%	-1.7%	-1.7%	-1.8%	3.5%	4.7%
UK	0.8%	1.1%	0.0%	-0.8%	0.2%	-0.6%	1.9%	1.5%	-1.9%	-1.8%	-1.3%	1.0%
Denmark	-4.5%	-0.4%	0.7%	0.4%	-0.1%	1.9%	-0.3%	-0.6%	-1.9%	-0.4%	0.0%	5.2%
France	-0.7%	0.9%	0.6%	-2.0%	0.1%	-1.3%	0.2%	0.5%	3.6%	-0.3%	-0.3%	-1.3%
Hungary	-2.4%	-0.9%	-1.3%	-0.3%	-0.1%	1.5%	-2.5%	0.0%	0.1%	0.5%	2.5%	2.8%
Portugal	-0.5%	-2.5%	1.4%	1.8%	-1.0%	-1.2%	0.7%	0.1%	1.0%	0.5%	1.1%	-1.5%

Table A10: Differences in the weighted proportional distribution of gender/age between 4th EQLS and ESS Round 7

	Male						Female					
	18-29	30-39	40-49	50-59	60-69	70+	18-29	30-39	40-49	50-59	60-69	70+
Austria	0.8%	-1.6%	0.4%	0.4%	-0.2%	0.2%	0.5%	-2.1%	0.0%	0.5%	0.6%	0.5%
Belgium	0.4%	1.1%	0.9%	-0.1%	-1.5%	0.1%	-0.3%	0.9%	-0.6%	-0.6%	-1.6%	1.2%
Czech Republic	-0.3%	1.9%	-1.4%	-0.7%	0.0%	1.0%	-0.4%	0.2%	-0.8%	-1.9%	-0.6%	2.9%
Estonia	0.0%	0.4%	-0.2%	0.0%	0.1%	-0.1%	0.4%	-1.2%	0.4%	-0.6%	-0.3%	1.0%
Finland	0.1%	1.3%	-0.1%	-0.7%	-0.7%	0.1%	0.3%	-0.1%	0.2%	-0.1%	-1.0%	0.6%
Germany	0.0%	1.0%	0.1%	-1.0%	-1.0%	0.7%	-0.1%	1.0%	-0.3%	-0.4%	-0.7%	0.7%
Ireland	-2.4%	1.5%	1.3%	-0.6%	-0.2%	0.4%	-1.6%	1.4%	0.3%	-1.0%	-0.9%	1.8%
Lithuania	0.5%	1.0%	-0.6%	-1.2%	0.0%	0.5%	0.3%	-0.2%	-0.8%	-0.3%	-1.9%	2.6%
Netherlands	-0.4%	0.9%	0.2%	-0.8%	0.1%	0.0%	0.0%	-0.2%	0.5%	-2.0%	-0.6%	2.3%
Poland	-0.8%	1.3%	-0.4%	0.6%	-0.4%	-0.4%	0.0%	0.1%	-0.5%	0.5%	-0.7%	0.6%
Slovenia	-1.9%	1.5%	1.6%	-0.8%	-0.4%	-0.2%	-0.6%	0.7%	0.0%	-0.5%	-0.6%	1.1%
Spain	-0.4%	-0.1%	-0.4%	0.1%	0.2%	0.1%	-0.8%	0.3%	0.8%	-0.8%	0.0%	0.9%
Sweden	0.4%	0.8%	-0.9%	0.3%	-0.2%	-0.8%	0.1%	-0.1%	1.2%	-0.8%	0.0%	-0.1%
United Kingdom	0.7%	-0.3%	0.3%	-0.2%	0.4%	-0.6%	0.5%	-1.0%	0.0%	-0.2%	-1.3%	1.7%
Denmark	-0.7%	0.8%	0.3%	0.1%	0.1%	-0.7%	1.4%	-0.2%	-0.7%	-0.2%	-0.6%	0.3%
France	1.8%	-1.6%	0.0%	0.3%	-0.2%	0.1%	1.8%	-2.6%	-0.3%	-0.1%	-0.9%	1.8%
Hungary	0.9%	0.3%	-0.4%	0.5%	0.2%	-1.2%	-0.8%	1.1%	-0.5%	0.0%	0.3%	-0.5%
Portugal	0.1%	-0.2%	0.3%	0.5%	0.0%	-1.4%	-0.2%	1.1%	0.0%	0.6%	-1.0%	0.3%

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

Table A11: Differences in the unweighted proportional distribution of household size and employment status between 4th EQLS and ESS Round 7

	Household size					Employment status
	1	2	3	4+		Employed
Austria	-4.4%	0.3%	2.6%	1.5%		11.0%
Belgium	19.9%	1.0%	-7.0%	-13.9%		-7.5%
Czech Republic	8.8%	5.8%	-8.7%	-5.9%		3.0%
Estonia	6.4%	1.1%	-3.4%	-4.1%		-3.5%
Finland	7.4%	1.6%	-4.8%	-4.2%		-3.5%
Germany	14.9%	-1.9%	-2.8%	-10.2%		6.5%
Ireland	-0.8%	2.4%	-1.6%	-0.1%		0.1%
Lithuania	15.2%	-0.9%	-8.7%	-5.6%		-3.6%
Netherlands	6.7%	0.1%	-3.5%	-3.3%		0.5%
Poland	3.9%	9.2%	-0.9%	-12.1%		-3.3%
Slovenia	1.2%	3.8%	1.5%	-6.5%		3.6%
Spain	4.6%	6.9%	-2.1%	-9.4%		0.0%
Sweden	0.2%	9.3%	-2.6%	-6.9%		-3.0%
UK	-3.3%	2.7%	0.7%	-0.1%		1.6%
Denmark	13.4%	-1.1%	-1.8%	-10.5%		-2.2%
France	5.0%	-4.2%	0.2%	-0.9%		2.8%
Hungary	15.3%	-3.7%	-6.8%	-4.8%		-2.5%
Portugal	3.5%	1.0%	0.3%	-4.8%		8.0%

Table A12: Differences in the weighted proportional distribution of household size and employment status between 4th EQLS and *ESS Round 7*

	Household size					Employment status
	1	2	3	4+		Employed
Austria	-10.0%	-1.2%	4.0%	7.2%		-1.7%
Belgium	4.3%	-0.9%	0.3%	-3.7%		1.7%
Czech Republic	2.4%	-4.3%	-4.3%	6.3%		-0.6%
Germany	6.4%	-2.1%	1.2%	-5.4%		1.2%
Estonia	-1.2%	-3.7%	-0.4%	5.4%		-0.3%
Spain	3.7%	-1.4%	0.5%	-2.8%		-2.7%
Finland	0.2%	0.2%	-1.8%	1.4%		1.1%
United Kingdom	-3.4%	4.6%	1.8%	-2.9%		3.9%
Ireland	0.0%	0.6%	0.3%	-1.0%		1.6%
Lithuania	6.5%	-5.4%	-7.4%	6.3%		0.7%
Netherlands	7.0%	2.1%	-2.4%	-6.6%		0.9%
Poland	0.4%	-2.6%	-1.7%	3.9%		-0.3%
Sweden	2.6%	2.2%	-1.8%	-3.1%		5.8%
Slovenia	2.9%	-0.3%	0.9%	-3.5%		3.4%
Denmark	6.5%	-1.2%	-0.4%	-4.9%		3.8%
France	7.1%	5.7%	-1.1%	-11.7%		4.5%
Hungary	-2.2%	-5.0%	-1.7%	8.9%		-1.2%
Portugal	0.0%	-0.8%	3.3%	-2.5%		2.0%

Annex 5. Paradata

Table A.13 Paradata Collected in the 4th EQLS
GPS coordinates (enumeration stage)
Final result code
Number of contact attempts
Interviewer ID, outcome, mode, and time for each contact attempt
Outcome and time for each CATI contact attempt
Household size
Sex of respondent/household member who refused or whose language/physical problems prevented interview
Age group of respondent/household member who refused or whose language physical problems prevented interview
Reason given by respondent/household member for not participating
Incentive received
Number of persons present during the interview, including interviewer
Respondent’s cooperation
Duration of interview
Interview back-checked
Interviewer gender
Interviewer age
Interviewer education

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency, whose role is to provide knowledge in the area of social, employment and work-related policies. Eurofound was established in 1975 by Council Regulation (EEC) No. 1365/75, to contribute to the planning and design of better living and working conditions in Europe.

Disclaimer: This working paper has not been subject to the full Eurofound evaluation, editorial and publication process.