WORKING PAPER

Eurofound

Working conditions and sustainable work

European Working Conditions Telephone Survey 2021: Data quality assessment

Working conditions in the time of COVID-19: Implications for the future

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

Authors: Javier Lorenzo Rodriguez, María del Mar Rueda García, Beatrice Cobo Rodríguez, Ramón Ferri-García, Lorenzo Navarrete Moreno, Laura Díaz Chorne, Marta Carretero Navarro and Victor Suárez-Lledo (Colegio Profesional de Politólogos y Sociólogos de Madrid, COLPOLSOC)

Research managers: Sophia MacGoris, Christopher White and Agnès Parent-Thirion

Eurofound reference number: WPEF24067

Related reports: *European Working Conditions Telephone Survey 2021: Quality assurance and control report* and *European Working Conditions Telephone Survey 2021: Technical report*

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency established in 1975. Its role is to provide knowledge in the area of social, employment and work-related policies according to Regulation (EU) 2019/127.

European Foundation for the Improvement of Living and Working Conditions

Telephone: (+353 1) 204 31 00

Email: information@eurofound.europa.eu

Web: www.eurofound.europa.eu

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

Abstract

This report presents the results of an independent quality assessment of the processes and outputs of the European Working Conditions Telephone Survey (EWCTS 2021) - an extraordinary edition, and the first telephone survey in the EWCS series due to the COVID-19 pandemic. This external evaluation assesses the quality of the survey processes from questionnaire design to fieldwork against Eurofound’s Quality Assurance Plan indicators, linked to the European Statistical System Quality Framework, that entails Relevance and Timeliness, Accuracy, Punctuality, Accessibility, and Coherence and Comparability. Survey outputs are then assessed against recent methodological literature and other comparable multinational, multiregional, and multicultural (‘3MC’) surveys from the European context. Based on the comprehensive review of the processes and outputs, our assessment concludes that the EWCTS 2021 has followed current best practices for ‘3MC’ surveys. Considering the critical nature of an assessment, recommendations are provided for improving the survey in the future.

Contents

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

List of tables

List of figures

Glossary of acronyms

'3MC': Multinational, Multiregional, and Multicultural Contexts

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

AAPOR: American Association for Public Opinion Research
ASA: American Statistical Association
CAPI: Computer-Assisted Personal Interviewing
CATI: Computer-Assisted Telephone Interviewing
CAWI: Computer-Assisted Web Interviewing
CCSG: Cross-Cultural Survey Guidelines
CPS: Professional Association of Political Scientists and Sociologists of Madrid
ESS: European Social Survey
ESOMAR: Market Research Association
EQLS: European Quality of Life Survey
EWCS: European Working Conditions Survey
EWCTS: European Working Conditions Telephone Survey
FAIR: Findable, Accessible, Interoperable, and Replicable
HHFA: Harmonized Health Facility Assessment
ISCED: International Standard Classification of Education
ISCO: International Standard Classification of Occupations
ISSP: International Social Survey Programme
ITC: International Test Commission
LASSO: Least absolute shrinkage and selection operator
LFS: Labour Force Survey
NACE: Statistical Classification of Economic Activities in the European Community
OECD: Organisation for Economic Co-operation and Development
PRO: Patient-reported outcomes
QAP: Quality Assurance Plan
RDD: Random Digit Dialling
SRSWOR: Simple Random Sampling Without Replacement
TRAPD: Translation, Review, Adjudication, Pretesting, and Documentation
TSE: Total Survey Error
WAPOR: World Association for Public Opinion Research
WHO: World Health Organization

Executive summary

Purpose

The European Working Conditions Telephone Survey (EWCTS 2021) 2021 surveyed 71,758 workers, between March and November 2021, across 36 countries including EU Member States, the United Kingdom, Norway, Switzerland, Albania, Bosnia and Herzegovina, Kosovo, Montenegro, North Macedonia, and Serbia. The EWCTS 2021 was an extraordinary telephone survey edition, carried out during the COVID-19 pandemic, after the European Agency for the Improvement of Living and Working Conditions (Eurofound)[1] had to terminate the face-to-face fieldwork for the 7th European Working Conditions Survey (EWCS 2020) after only seven weeks. This specific context required switching to a safe data collection mode which would be compatible with mobility restrictions, using random digit dialling (RDD) of mobile phones. Other adaptations followed this decision, such as the necessary shortening of the questionnaire performed through its modularisation, or adaptations in the response scales. The result is a unique survey in the EWCS series that provides a wide-ranging picture of job quality across countries, occupations, sectors, gender, and age groups in the context of the COVID-19 pandemic.

A multidisciplinary research team of sociologists, political scientists, statisticians, and survey methodologists, specialized in survey design and survey methodology, led by The Professional Association of Political Scientists and Sociologists of Madrid (CPS), was awarded the contract to conduct an independent quality assessment of the processes and outputs of the EWCTS 2021 and to provide some recommendations for improving future editions. The report details the results of this quality assessment.

Methodology

Building on Eurofound’s previous quality assessments and current best practices and literature on multinational, multiregional, or multicultural surveys (referred to as ‘‘3MC’’ surveys), this report combines state-of-the-art methodological approaches to survey quality assessment in an integrated framework, specifically: monitoring survey production process quality, fitness for intended use, and total survey error.

The quality assessment of the survey processes is organised along the main stages of the survey lifecycle: questionnaire development, adaptation and translation, sampling, and weighting (although these two enter the domain of outputs too) and fieldwork and carried out against Eurofound’s Quality Assurance Plan (QAP). The QAP is a comprehensive set of quality indicators and associated targets, linked to quality dimensions determined in the European Statistical System as a framework to assess quality in terms of Relevance and Timeliness, Accuracy, Accessibility, Coherence and Comparability, and Punctuality.

[1] The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency established in 1975. Its role is to provide knowledge to assist in the development of better social, employment and work-related policies according to Regulation (EU) 2019/127.

The quality assessment of outputs, entailed the review of paradata and microdata, considering among other aspects, internal and external validity. An evaluation performed against current best practices for '3MC' surveys, such as the European Social Survey (ESS) or the Labour Force Survey (LFS), and recent methodological literature, such as the Cross-Cultural Survey Guidelines (Survey Research Center, 2016) or AAPOR WAPOR Task Force Report on Quality in Comparative Surveys (2021).

Given the exceptional circumstances in which the survey was conducted, the evaluation examines all survey processes and phases, with special attention to: the transition from Computer-Assisted Personal Interviewing (CAPI) to Computer-Assisted Telephone Interviewing (CATI) and the implications for future transition to Computer-Assisted Web Interviewing (CAWI); the impact on the survey outcomes of the contingencies made due to COVID-19 (Modularization, Re-scaling, Translation, etc.); the role and impact of the Quality Assurance Framework and the Quality Assurance Plan (QAP); the comparability of the data at cross-country level, with the time series' EWCS data, and with other international surveys; along with the complexities of documenting all of it in a comprehensive manner with an attractive and accessible dissemination of its results.

The process of assessment followed a triangulation methodology that allowed for an all-encompassing and multi-sources information flow (Flick, 2002). For quality assessment, mixed methods offer a depth of qualitative understanding with the reach of quantitative techniques (Jahoda et al., 1976; Fielding and Fielding, 1986; Denzin, 2010). This methodological strategy fits with the available diversity of quantitative and qualitative sources and the vast array of documentation produced by Eurofound and Ipsos, the fieldwork contractor that conducted the CATI survey on behalf of Eurofound.

Key findings

Overall, the evaluation of the quality of the survey processes is deemed as having a high level of compliance with Eurofound's QAP indicators across all stages. The QAP has served as a relevant, robust, and comprehensive tool to track and control the quality of all processes along the survey lifecycle. This is particularly notable given the pressing circumstances and the change in administration mode. The quick adaptation of the QAP to serve the purpose of a telephone survey was swift and granted a quality process. There were some minor deviations, but this non-compliance or almost-compliance in some cases, is assessed as having a minimal effect on data quality.

The QAP itself is generally assessed as a great framework against which to monitor the quality of the survey processes. Although this report suggests alternative indicators, those already in place are generally relevant, appropriate, and comprehensive. If anything, the list could be reduced or optimized to facilitate the work of the fieldwork contractor, signalling those that are more relevant for the overall quality of the outputs.

The transition to the EWCTS 2021 must be considered as a reference of rigour, professionalism, and determination to do the best possible work in perhaps the worst conditions. Overall, it can be stated that the processes carried out in the EWCTS 2021 extraordinary edition are (in most cases) up to best practices and standards on '3MC' surveys.

The questionnaire development process incorporated many current best practices such as consultation with subject matter experts and stakeholders, an advance translation of the questionnaire including the triangulation of two expert linguists in survey translation and cross cultural survey translation (‘3MC’), overall assessment by an expert in ‘3MC’ survey methodology, fully translated following a simplified TRAPD (translation, review, adjudication, pre-test, documentation) approach, harmonisation and adaptation, and a sizable investment in training, cognitive and piloting pretesting. The EWCTS 2021 questionnaire was developed by adapting the previously designed and tested face-to-face questionnaire to suit a telephone interview. This process involved shortening the questionnaire, reducing response scales, and adapting and shortening the introduction and final questions via a modularised approach. This questionnaire made use of the advance translation, cognitive test and TRAPD translation previously produced for the CAPI questionnaire; the new questionnaire, adapted for CATI, which was very similar to the CAPI one, was additionally subjected to a simplified TRAPD approach (with one translator and one adjudicator), harmonisation and adaptation processes and fully tested in a pilot in all countries. Overall, the processes retain their quality and included current best standards of ‘3MC’ questionnaire design.

Despite the need for different sample designs due to changing conditions, the survey successfully adjusted its sample calibration and treatment procedures accordingly. Overall, EWCTS 2021 followed sound principles for its sampling design and weighting procedure, ensuring its comparability across all the participant countries, with sample sizes being large enough to produce reliable national estimates and the fieldwork took place without significant issues. The weighting system was implemented following regular standards used in calibration, with a proactive construction of design weights taking over coverage into account, a calibration procedure in various steps to avoid further problems, using auxiliary variables that may have correlations with potential variables of interest, and using linear bounded distances which avoids further weight trimming. In addition, the analysis of the weighting procedure was thoroughly documented in the Sampling Report.

The fieldwork process was meticulously planned and closely monitored, allowing for the prompt detection and resolution of issues. Weighting adjustments have been applied to minimize nonresponse bias, a common challenge in telephone surveys, by utilising available auxiliary variables. Nonresponse has become an important issue in probabilistic surveys, especially with declining participation rates, particularly evident in telephone surveys (Beullens et al., 2018). This survey is no exception, and a notable problem is the very high non-response rate. It should be noted however that this is, as discussed, a common issue in CATI surveys also experienced by other big surveys developed during the pandemic, such as the Labour Force Survey or Americas Barometer (Hox & De Leeuw, 1994; De Rada, 2015, AAPOR, 2021; Castorena et al., 2022; Eurostat, 2022) and that the alternative during the pandemic would have been a non-probabilistic survey. It is important to note that the non-response rate alone does not directly indicate non-response bias for a specific survey, and in this case the percentage of non-response is quite balanced in the categories of the sociodemographic variables considered, so the reweighting methods used were useful to reduce the observed bias.

Overall, the use of a standardised CATI instrument in the EWCTS 2021 facilitated a standardised collection of paradata. In relation to the microdata, the quality assessment and comparison to previous rounds was difficult given the change in the mode of administration, the questionnaire adaptation to interviews conducted over the phone instead of face-to-face, which entailed the use of different (reduced) response scales, and the inclusion and exclusion of questions. In addition, changes in the results could be due to real changes in working conditions during COVID-19. In terms of survey quality, there were no relevant issues found that could have seriously compromised the internal and external validity, reliability, or overall quality of the survey results.

The recommendations mainly focus on enhancing the survey's efficiency, effectiveness, and accuracy, which are expected standards for a high-quality survey like this one. Despite the challenges highlighted in the report regarding the survey's execution, and particularly the effects derived from the change of administration mode regarding country coverage, nonresponse, and comparability, no major issues with a relevant impact on quality were detected.

Recommendations

These recommendations have been carefully devised through a method that involves a convergence process, including expert interviewees and the evaluation research team:

Quality Assurance:

- Optimise the Quality Assurance Plan.
- Ensure a consistent labelling and workflow both at back office and front office survey processes.
- Expand information on the decision making in the reports.

Questionnaire Design:

- Develop an analysis plan to reduce the questionnaire.
- Continue and develop the Glossary and Concordance Grid.
- Ensure specific demographics remain engaged.
- Implement Methodological Workshops.

Cognitive Testing (CT):

- Set standards on the methodology and reporting of cognitive tests.
- Adapt cognitive test to new administration modes.
- Consider web probing in the next survey.
- Include additional variables in the CT sample.

Translation:

- Ensure a team approach in the review and adjudication.
- Continue to ensure the pretesting of all languages.

Sampling and Weighting:

- Implementing new variance estimators
- Using of multiple frame estimators
- Selecting variables for calibration by country or propensity score adjustments
- Include details from the procedure to allocate sample size in the sampling report.
- Leave adjustment uncapped if CATI is used again.

Fieldwork:

- Reduce respondents’ language barriers.
- Develop more visual or interactive training materials.
- Exclude willingness to be recontacted from interview time calculations.

- Control implementation of planned missingness designs

Microdata and paradata

- Include information on the original weights of the design in the data file.
- Develop an advanced response analysis from all participating countries.
- Refine data variable information.

Reporting and dissemination

- Harmonisation between Eurofound´s website and the UK Data Archive.
- Harmonise nomenclatures
- Use permanent links or redirections
- Enhance the user´s experience.
- Update and make public the Concordance Grid and Glossary.
- Consider new access routes to the UK Data Archive or other data repositories

1. Introduction

1.1. Background of the EWCTS 2021 extraordinary edition

Since its launch in 1990, the European Working Conditions Survey (EWCS) has provided an overview of working conditions in Europe. The survey collects unique and critical data on both employees and the self-employed across Europe on a harmonised basis. Its scope and geographical coverage have widened substantially since the first edition, aiming to provide a comprehensive picture of the everyday reality of women and men at work on a broad set of issues such as employment status, working time and organisation, learning and training, physical and psychosocial wellbeing, risk factors, work-life balance, worker participation, earnings and financial security, groups at risk, etc.

This report is an external quality assessment of the European Working Conditions Telephone Survey (EWCTS 2021), which surveyed 71,758 workers, between March and November 2021, across 36 countries including all of the EU Member States, the United Kingdom, Norway, Switzerland, Albania, Bosnia and Herzegovina, Kosovo, Montenegro, North Macedonia, and Serbia. The 2021 EWCTS 2021 was an extraordinary telephone survey edition, carried out during the COVID-19 pandemic, after the European Agency for the Improvement of Living and Working Conditions (Eurofound) had to terminate face-to-face fieldwork in March 2020 for the 7th European Working Conditions Survey (EWCS) after only seven weeks due to the pandemic. The Agency's prompt and swift response and efforts granted the continuation and adaptation of the survey, allowing for the study and monitoring of such issues at a particularly relevant and challenging time for employment and working conditions, while still ensuring its quality control and assurance. This specific context required switching to a safe and data collection mode compatible with mobility restrictions, by using random digit dialling (RDD) of mobile phones. Other adaptations followed this decision, such as the necessary shortening of the questionnaire performed through its modularisation, or adaptations in the response scales. The result is a unique survey in the EWCS series that provides a wide-ranging picture of job quality across countries, occupations, sectors, gender, and age groups in the context of the COVID-19 pandemic.

In general, the EWCS has substantial relevance and serves as a pivotal resource, impacting various sectors and stakeholders. Its significance stems from the reliable and comprehensive insights it provides on the ever-evolving landscape of work and employment across the European region. Traditionally, in each wave a probabilistic random sample of workers, employees and self-employed has been interviewed face-to-face. The survey series, carried out every four to five years, is comparative by design and replicate a significant number of questions from wave to wave, allowing not only cross-country comparison but to monitor trends by providing homogeneous indicators on employment and quality of work at national and European level, contributing to policy development.

This continuity provides a unique perspective on how working conditions, employment patterns, and job quality have evolved in response to economic, technological, and societal changes. By offering a comprehensive understanding of factors such as employment patterns, job quality, and working conditions, the survey aids in deciphering the intricate interplay between economic trends and the workforce, providing insights into emerging trends such as remote work, gig economy engagements, and changing work arrangements. It also applies a person-centred perspective and holistic view of job quality, delving into aspects beyond employment conditions, such as work-life balance, psychological well-being, and career development. A comprehensive approach that aids in evaluating the overall impact of work on individuals' lives.

In summary, the EWCS stands as a cornerstone fostering high quality cross-country comparisons, tracking trends, supporting academic endeavours, facilitating meaningful dialogues, monitoring well-being, and informing policymakers and anchoring decision-making in solid evidence.

Accordingly, Eurofound has consistently emphasized the importance of data quality, both internally and externally, in all editions of the EWCS. Besides the in-house quality controls conducted by both Eurofound and the fieldwork's contractor (Ipsos), Eurofound has requested an external and ex-post comprehensive quality assessment. The Professional Association of Political Scientists and Sociologists of Madrid (CPS) was awarded the contract to carry out this independent assessment. This report presents the results and recommendations of the assessment.

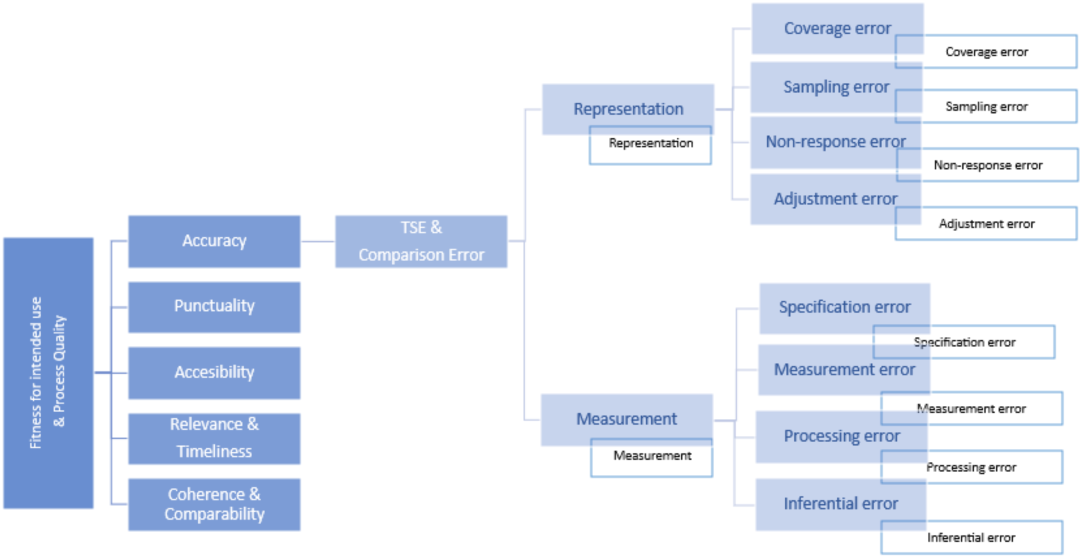
1.2. Methodological approach to the EWCTS 2021 quality assessment

Following Eurofound's tender criteria, the evaluation examines all survey processes and phases, with special attention to:

- The transition from Computer-Assisted Personal Interviewing (CAPI) to Computer-Assisted Telephone Interviewing (CATI) and the implications for future transition to Computer-Assisted Web Interviewing (CAWI)
- The comparability of the data with the EWCS time series, with other international surveys and at cross-country level
- The impact on the survey outcomes of the contingencies made due to COVID-19 (Modularisation, Re-scaling, Translation, etc.)
- The role and impact of the Quality Assurance Framework and Quality Assurance Plan, along with suggestions for improvement

To this end and building up on previous external assessments, particularly the EQLS 2016 Quality Assessment, different approaches to survey quality assessment are used in an integrated framework, specifically: monitoring survey production process quality, (Biemer & Lyberg, 2003; Gryna, 2001; Groves & Heeringa 2006), fitness for intended use, and Total Survey Error or TSE (Groves, 2009; Biemer, 2010, 2016; Groves & Lyberg, 2010; Lyberg & Weisberg, 2016).

Figure 1. Integrated quality framework



Source: Author's own elaboration based on Eurofound (2021), Pennell et al. (2017); Smith, (2011)

The first two are observed in the assessment of the quality of survey processes organised along the main stages of the survey lifecycle: questionnaire development, adaptation and translation, sampling, and weighting (both sampling and weighting are included here but include considerations related to the outputs, such as the evaluation of the final composition or the calibration and poststratification of the outputs) and fieldwork. This assessment is carried out against Eurofound's Quality Assurance Plan, which comprises a set of indicators related to the quality dimensions of the European Statistical System framework:

Table 1. Eurofound's quality criteria

Criteria	Indicator
Relevance & Timeliness	Relevance for users of the survey data and survey-based reports, both in terms of substance and timing of publication
Accuracy	Validity and reliability of the survey data
Accessibility	Availability of outputs and transparency of processes
Coherence & Comparability	Consistency with other datasets as well as internal comparability (e.g., comparability between countries or groups within the survey)
Punctuality	Adherence to the timeline as set at the start of the project.

Source: Eurofound (2021): EWCTS 2021 Quality assurance and control report.

The assessment of the QAP indicators is provided both in terms of their achievement, providing an external evaluation of their completion independent to that of Eurofound and Ipsos, and in terms of a general reflection on their relevance, appropriateness, and comprehensiveness. In regard to the first of these objectives, indicators are marked as either "met," "not met," and on some occasions as "not applicable" in specific cases where the indicator was not relevant for the CATI mode of administration, or “mostly met”. The latter applied in specific cases where the slightly deviated from the deadline, or the pre-established goals (e.g. reach 39 out of 40 interviews). That is, when very minimal and not relevant for quality deviations from the target occurred.

Unlike some previous assessments that focus solely on the indicators related to the accuracy dimension, the aim is to address the other quality dimensions as well. While the capital relevance of accuracy for all other dimensions (Biemer & Lyberg, 2003) is recognised, it could be equally argued that an extremely accurate survey becomes inapt if not relevant, comparable, published on time or accessible to the users (Madans et al., 2011).

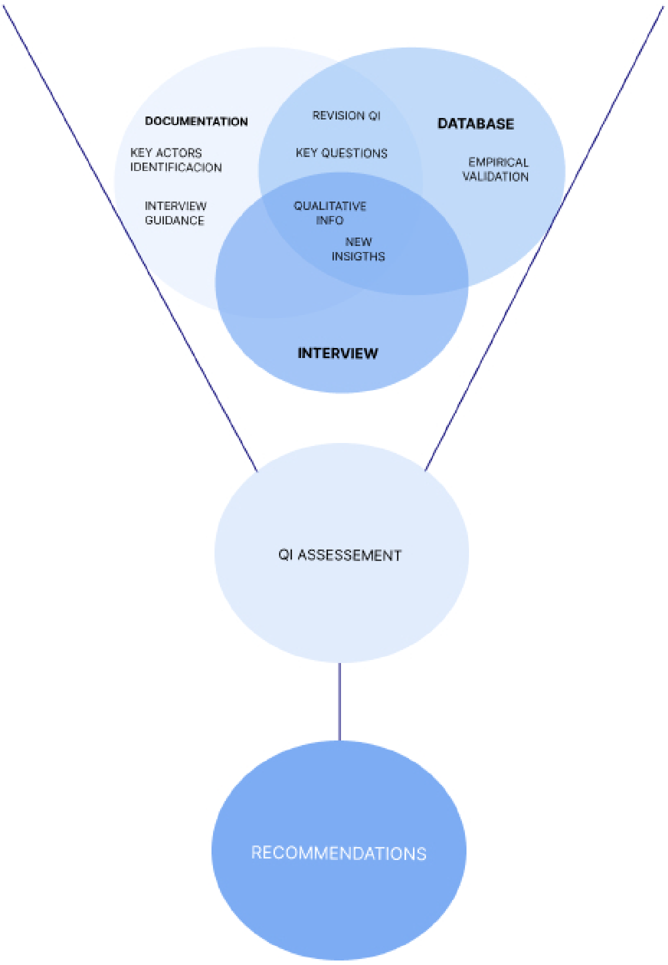
The approach in the assessment of the quality of the outputs, which entails the review of paradata and microdata quality, considering among other things: nonresponse, internal and external validity, is set in the framework of the Total Survey Error approach (Weisberg, 2009; Groves and Lyberg, 2010; Biemer, 2010), following state-of-the-art methodological literature, and current best practices for ‘3MC’ surveys, such as the European Social Survey (ESS), Labour Force Survey (LFS), the Cross-Cultural Survey Guidelines (Survey Research Centre, 2016) or AAPOR WAPOR Task Force Report on Quality in Comparative Surveys (2021).

Total Survey Error is the dominant paradigm in the field of survey methodology which aims to identify all sources of bias (systematic error) and variance (random error) that may affect the validity (accuracy) of survey data, assisting to evaluate design and implementation trade-offs maximising data quality (Biemer, 2010). There are several TSE typologies as different scholars include different source of error (Groves & Lyberg, 2010). Total survey error (TSE) defines quality as the estimation and reduction of the mean square error (MSE), which is the sum of random errors (variance) and squared systematic errors (bias) both in measurement (specification error, measurement error, and processing error) as well as representativeness (coverage error, sampling error, non-response error and adjustment error). Although some of these elements, like the sampling variance, can be measured in most probability sample surveys, others cannot without significant design burdens. So many times, TSE is used to define a quality approach in which attention is devoted to the entire set of survey design components (Groves & Lyberg, 2020). Applied in a comparative ‘3MC’ context and together with comparison error, this approach informs design and implementation trade-offs while maximising comparability or equivalence (Weisberg, 2005; Pennell et al., 2017; Słomczyński 2019; Smith, 2011, 2019, Roberts, 2020; AAPOR, 2021). The ultimate ambition being to continually improve the processes and optimally allocate the resources to minimise critical or more relevant errors while maximising comparability or equivalence.

The end goal of the external assessment is to examine the processes and value of the QAP indicators: whether they are adequate, specific, and correctly measured. In addition it aims to provide an independent evaluation of the overall quality plan and framework, adding insightful recommendations. It is worth noting that the assessment of quality is an exercise that necessarily entails pointing at issues that could be further improved and to provide recommendations to enhance the survey quality in the future. This should not be mistaken for a general criticism of the EWCTS 2021 overall quality, which after careful consideration, and especially given constraining circumstances, we assess as exceptional in many aspects and comparable to the best current standards. Many of the issues encountered are common to other ‘3MC’ surveys and either difficult to address in terms of cost/benefits or a direct consequence of the exceptional mode of administration (CATI). Others are, however, easier to address, recurrent or hold a higher importance for the quality of the series overall. The quality enhancement of a survey is a never-ending process, encompassing a constant learning and improving procedures, as proved over the years by the development of the QAP, the advance translation and cognitive tests, the development of technical reports with the motivation of the new questions added in the questionnaire, etc.

The assessment process demanded a triangulation methodology that allowed for an all-encompassing and multi-sources information flow. The triangulation worked as an integrated approach for gathering data, having a 360-theory approach (Flick, 2002). Triangulation methods have been used in Social Sciences and Public Policy Evaluation since Campbell & Fiske (1959). For quality assessment, mixed methods offer a depth of qualitative understanding with the reach of quantitative techniques. The combinatory use of both has been used by scholars in social sciences (Jahoda et al., 1976; Fielding and Fielding, 1986; Denzin, 2010). This methodological strategy fits with the diversity of information sources available and the vast array of documentation produced by Eurofound. Both primary and secondary and qualitative and quantitate sources of information were used. The documentation source consisted of the review of extensive qualitative and descriptive documentation related to each survey process gathered and produced by Eurofound and Ipsos such as the following reports produced for the EWCTS 2021: Translation report; Data validation and editing report; Sampling & Weighting report (Eurofound, 2021a, 2021b, 2021c); Quality Assurance and Control Report; Pilot Report; Technical Report; Sampling and Weighting reports (Ipsos, 2021a; 2021b; 2021c; 2021d), etc. This was enriched with contributions from the latest and relevant literature. A second information source was the main database, and the analyses of the microdata and paradata. These sources of information were completed with fourteen qualitative interviews (detailed in table 41) with expert actors with relevant experience in the distinct aspects of the survey or directly involved in the implementation or design of the EWCTS 2021. To offer the most feasible recommendations for Eurofound's future EWCS rounds, the methodology proposed implies a convergence process (Thomas, et al., 2021; Van Praag, 2021) by which key actors discriminated and prioritised the recommendations.

Figure 2. Methodological process and workflow



Source: Author's own elaboration

The structure of this report follows the logical structure of the survey lifecycle, beginning with the quality assessment of the survey processes from questionnaire planning and design to fieldwork, sampling, and weighting process. The second section of this report asses the quality of the survey outputs, the microdata and the paradata and the documentation and dissemination of survey outputs. Every section keeps the same structure, beginning with a brief introduction and the specific methodological particularities associated with the assessment of that task; the indicators included in the evaluation of this task and their assessment. Lastly, there is a final section of conclusions and recommendations. At the end of the document, a complete list of references is provided, as are the support analyses in the Appendix.

2. Survey processes quality assessment

This section contains a comprehensive evaluation of the questionnaire design, cognitive testing and translation, fieldwork, sampling, and weighting processes. Given the complexity and quantity of processes involved in the survey, each of these topics has its own subsection. Thus, the first subsection reviews the questionnaire planning and design, encompassing the questionnaire development and adaptation to CATI, cognitive testing, and the translation process. The second part assess the fieldwork process, while the final subsection deals with the quality assessment of the sampling and last, weighing processes.

2.1 Questionnaire design and translation

Assessing the quality of a long-lasting questionnaire such as the EWCTS 2021 is a challenging endeavour. On the one hand, its quality is well-proven and thoroughly evaluated in previous editions. On the other hand, this edition presents several changes that merit an in-depth evaluation. The quality assessment of these processes has also been more complex not only due to the change of methodology but also to the fact that the questionnaire development has benefited from tasks previously carried for the CAPI phase. This multiplies the documents of reference and sometimes makes it difficult to trace decisions or assess their consequences.

This section addresses the quality assessment of the questionnaire design and its adaptation from CAPI to CATI, cognitive testing, and translation, harmonisation, and adaptation. Following the methodological strategy already described, the assessment is based on a thorough analysis of the QAP indicators, after exhaustively reviewing the vast array of documentation produced by Eurofound (see the References section) and the qualitative information obtained through interviews with experts and relevant actors.

The primary objective when designing ‘‘3MC’’ questionnaires is to ensure survey questions are comparable and equivalent across languages and cultures while minimising specification and measurement errors related to questionnaire design. This section discusses some of the difficulties of this endeavour and emphasizes that in a ‘3MC’ context, questionnaire design cannot be detached from translation and adaptation. Finding and effectively managing a team with the necessary expertise and knowledge can be very complex. Additionally, the challenges of documentation, quality assurance, monitoring, and assessment for questionnaire design are also more complex in a ‘3MC’ context (Harkness et al., 2010).

Further development of theory and research on the influence of culture and cognition on survey response is relevant for advancing ‘3MC’ questionnaire design. While initial theories have integrated culture into survey response models, the complexity of the emerging picture calls for ongoing theoretical debates among cultural psychologists regarding the dimensions and conceptualization of culture and the extent to which culture can be viewed as an explanatory variable. Research demonstrates that cultural mindsets can be activated based on the situation, leading to different perceptions and behaviours (Pennell et al., 2017). Thus, it is important to consider how differences in the response process may be influenced in the moment by various aspects of the research context.

Establishing cross-cultural validity is vital for questionnaires designed to compare data. However, it is common practice to avoid testing measurement equivalence, or equivalence is only tested for a limited selection of questionnaire items. Additionally, pretesting multiple alternative measures can be time-consuming and costly, with limited evidence available to guide decision-making. All these guidelines for achieving a ‘‘3MC’’ approach can be affected when transitioning from one survey mode to another, where the methodology may be impacted, and the harmonisation process can be more costly. In this regard, Eurofound has sought to implement and maintain the essence of the ‘3MC’ approach surveys when transitioning from the CAPI to CATI, aiming to achieve the highest comparability in the Questionnaire Design process. To assess the good practices developed in this process, it is necessary to compare them to other processes evaluated using the TSE and the ‘3MC’ approach, to establish recommendations for future waves (Hox and Leeuw, 1994).

2.1.1 Questionnaire design and planning

2.1.1.1 Analysis of quality indicators on questionnaire design

Most quality indicators contained in the Quality Assurance Plan (QAP) related to questionnaire design were successfully met.

Table 2. Questionnaire design quality indicators

Number	Criteria	Indicator	Target	Assessment
57	Relevance & Timeliness	Questionnaire has been consulted with Eurofound's stakeholders/Advisory Committee	Y	Target met
58	Punctuality	Timeline for questionnaire development is kept	Y	Target met
59	Accuracy	Comprehensive Glossary is provided on time	Y	Target mostly met

Source: Author’s own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

These three indicators have been consistently used in previous EWCS editions and in the European Company Survey 2019 (Desiree & Lenaerts, 2020). This suggests that the procedures employed for the development of the surveys have remained consistent and comparable across different data collection efforts, which helps ensure the continuity and reliability of the data obtained. Although the indicators were primarily designed and evaluated for the EWCS CAPI phase, they have proven to remain relevant and useful despite the changes in data collection modalities.

The deadlines and consultations with stakeholders and other collaborators were carried out in line with the plans, and the quality of the survey and the questionnaire were not affected by the pressing circumstances. Evidence has been provided on the fact that the original questionnaire went through many reviews and consultations with stakeholders, policy users, researchers, Eurofound staff and potential respondents (Ipsos, 2021a) including:

- The Advisory Committee which involves representatives from Worker, Employer, Government and European Commissions groups on the Management board of Eurofound. 4 meetings (October 2017 to March 2019)
- The Expert Questionnaire Group which involves international users such as International Labour Organization (ILO), Organisation for Economic Co-operation and Development (OECD),

representatives from the European Commission and European Agencies such as the Statistical Office of the European Union (Eurostat) or the European Agency for Safety and Health at Work (EU-OSHA), representatives from national working conditions surveys, representatives from Eurofound tripartite stakeholders (the members of the Advisory Committee) and experts on working conditions topics from the various disciplines. 2 full group meetings (December 2017 and November 2018). 2 in-depth meetings explored issues identified by the main group: deepening and making the business case for job quality; workers at the margin and the fissured workplaces.

- Eurofound Activity Group “working conditions and sustainable work” ad hoc meetings between November 2017 and December 2018 were dedicated to the revision of the questionnaire.
- Eurofound consultation meeting involving all researchers from other Activity groups took place in December 2018.
- European Commission Directorate-General for Employment, Social Affairs, and Inclusion (DG EMPL) series of ad hoc and in-depth consultations.
- EWCS quarterly meetings; the Director, Deputy Director, Head of research, Head of the Unit responsible for the EWCS and the Project Manager met once every quarter to discuss progress and confirm some strategic discussions.
- Potential respondents were involved in the revision of the questionnaire through the cognitive pretest.
- The Advance Translation served also as a source to detect errors or issues in the questionnaire design early on.

Additionally, the new modularised version (EWCTS 2021 questionnaire) which took the EWCS questionnaire as basis, was also presented to EWCS stakeholders in an abbreviated process:

- The Advisory Committee Working Conditions meeting (September 2020)
- The Expert Questionnaire Group (September 2020)
- Presented to partners in Norway, Slovenia Belgium, and Switzerland.

The questionnaire timeline was kept and a comprehensive glossary, using the one developed for EWCS 2020, was delivered. The glossary was however delivered three days after the translations were initiated, additionally there are some discrepancies in the reviewed documentation (Translation Report, Quality Assurance and Control Report, Quality Plan) regarding the accomplishment of this indicator. Given the short delay, the impact on the overall quality of the survey is not deemed major, however given the tight deadlines of the translation processes all efforts should be made in the future to keep this deadline and preserve the glossary structure (more on this below).

The Glossary facilitated in the translation process to support the functionally equivalent translation of the key terms is considered a good quality practice that build on recommendations from previous assessments. So is the Questionnaire Concordance Grid (1999-2015), a tool relevant for any researcher, practitioner or those who make use of the survey (for example, Giménez-Nadal et al., 2022). The recommendation is to work on their continuation and refinement ensuring their timely update and availability.

Considerations from the glossary like the rationale for including or modifying question, the expert assessment and the source or international standard from which the question was taken or adapted could enhance not only the accuracy of the translations, but the use and comparability of the survey, so it is recommended to continue it and to make it fully or partially available to the public by, for example, including some parts of it in the concordance grid. Another recommendation is to continue and further develop the "measurement objectives" column of the Glossary on how the question has or will be exploited, or otherwise to develop an analysis plan, which could contribute to the shortening of the questionnaire or prioritization of questions.

In the assessment, at least two documents with the label glossary were evaluated (one for translators and other with Eurofound and an external expert) and additionally a Word document, also provided to translators and interviewers with similar information. Ipsos also report having to merge the 2015 and 2020 files and reordered the questions (Ipsos, 2021a, p. 11), which probably had an impact on its delivery time and potentially the quality of translations. Although the development of these tools and documents is certainly commended as a good practice, efforts should be made to normalise labels and formats and to condense the information in as few documents as possible. The use of a database could be considered.

2.1.1.2 Comparability of the questions in EWCS editions

EWCTS 2021 questionnaire was developed, as previously discussed, by adapting the previously designed and tested EWCS 2020 questionnaire. The main challenge was to adapt the questionnaire designed for a face-to-face methodology into one to be administered by a telephone interview (Ipsos, 2021c). This process involved shortening the questionnaire, adapting the questions which no longer had showcards or supporting materials, adjusting sensitive questions, reducing response scales, rearranging the question order, and adapting and reducing the introduction and final questions.

The reduction of the questionnaire proved to be a difficult endeavour given the relevance of the results for stakeholders and policy makers in the context of the pandemic. As a result, Eurofound adopted a planned missingness design via a modularised approach in which rather than cutting the questionnaire, which would have resulted in a loss of information, it was divided into three sections or modules, one mandatory for all respondents, randomly allocating the rest of the sample into the other two modules with six possible paths. The items included in the modules were, according to Eurofound (2022), those most relevant to workers' well-being and with the strongest evidence for their statistical reliability. Such a method is considered viable given the considerable increase in the sample for the EWCTS 2021. Although the comparability of the series was affected both by the mode of administration, and the sample responding to different modules of questions but not the entire questionnaire, it still allows for the tracking (with due caution) of most variables of interest. It is also an exercise relevant for survey methodologists to test modularisation to a level rarely undertaken for a '3MC' survey. A few challenges arose in the implementation and the automated allocation mechanism of this planned missingness design which are further detailed in the fieldwork and weighting section.

The questionnaire was adapted, and some questions were removed or shortened to reduce the duration of the questionnaire in the adaptation process to CATI. For example, the questions concerning household were simplified.. After conducting the Cognitive Test, Eurofound team also decided to modify, for a better understanding, some questions related to the classification of "self-employee/employee," with the objective of contributing to an analysis of "dependent self-employment." This addresses employees who are given the option of either being fired or starting to work as self-employed, only to be rehired on a private contract basis by their former employer.. Another update from the latest wave is the inclusion in the gender question with the option "Or would you describe yourself in another way?”. Overall, the questionnaire design followed best current practice and is well documented. Eurofound efforts to explain the rationale behind the modification, adaptation, or elimination of questions, is reflected as stated before in the Glossary which includes an external ‘3MC’ expert assessment of the questionnaire. The glossary used, was the same as the one designed for the evaluation of EWCS 2020. The reason behind using the same glossary was that it remained applicable for the latest edition, with only two new questions added related to COVID-19 vaccination guidelines.

In general, the questions that were introduced in the last wave did not pose difficulties in terms of comprehension for the interviewees, indicating that the questionnaire design successfully maintained its quality. The Fieldwork section of the Pilot report, includes some points in relation to the questionnaire design. In both, the sixth edition (2015) and the Pilot Reports of 2020 and 2021, most comments or complaints involved the questionnaire duration. The fact that the considerable reduction of the survey due to CATI and modularization did not significantly affect these complaints (although the different modes of administration make comparisons difficult) show that these might be, to some extent, structural complaints related to the size, scope, and difficult topic of the survey, which includes many technical terms and concepts which might be difficult to apprehend. These comments are not uncommon in comparative surveys, and it must be considered that there is always a trade-off in terms of quality, in which an improvement in the response rate, engagement of the participants, and accuracy would be made at the cost of the survey’s relevance. However, they should not be completely disregarded and efforts to continue reducing wherever possible the length and difficulty should be furthered. In some instances, the reductions have been proven to be straightforward and easy to make; for example, complaints about the introduction length and content during the pilot were swiftly addressed by completely re-writing the text to be more informal, concise and avoid terms with different or conflicting cultural meanings like “policy makers” or “personal data”. Other reductions present more difficulties, and potential solutions revolve around continuing the work on defining the rationale for including questions and further developing how they have been or should be exploited; continue testing the modularization of certain questions; or testing the duration of the survey across profiles and paths to find where the efforts should be increased. The EWCTS 2021 Pilot Report identifies that certain respondents experience longer durations and more difficulties, specifically the older workforce and respondents with lower levels of education, different paths for self-employees and employees, or language versions also affect the duration. It would be important to continue including these criteria in the cognitive testing sampling and include them in pilot analyses.

Overall, the evaluation of the questionnaire in the pilot and fieldwork by local agencies and Ipsos’ Central Coordination Team was positive. Despite some complaints regarding the length of the questionnaire, this did not adversely affect the quality of responses from the interviewees or their overall motivation towards the survey. Many issues detected during the pilot were considered and addressed by adapting or changing the questionnaire. Therefore, it can be concluded that the practice carried out by Eurofound has been good and has maintained the survey's quality.

2.1.2 Cognitive testing

Cognitive tests aim to assess the effectiveness of new and updated survey questions by evaluating how well respondents understand the terms and concepts and to analyse the answering and thought process of respondents and their ability to provide clear answers that meet questionnaire drafters expectations. Testing of survey questionnaires has long been recognised an important aspect of data quality assurance, which has been traditionally, and sometimes exclusively, reliant on expert reviews (Goerman et al., 2018) and field testing. Current best standards recognise that response errors can easily go unnoticed if survey questionnaires are not tested by their target population to assess the cognitive processes participants use to interpret and answer questions (Presser et al, 2004). In '3MC' contexts it can also be used to assess the questionnaire measurement equivalence and hence its comparability, helping detect cross-cultural variations in response, construct validity (Miller, 2019), and translation issues (Behr and Braun, 2015; Braun et al., 2018; Meitinger, 2017; AAPOR, 2021). This testing helps uncover difficulties in a lengthy technical survey like the EWCS.

As the CAPI and CATI questionnaire versions were similar and had already been thoroughly tested, and due to time and budgetary constraints the EWCTS 2021 questionnaire relied on the Advance Translation and Cognitive Test carried out for the CAPI version between February and April 2019, and no further testing took place, except for a full pilot in all countries for the EWCTS 2021.

The CAPI cognitive testing was carried out in two countries: Ireland, and Poland. Twenty interviews were conducted in each country, and the results were used to refine the survey protocol and the Glossary. Cognitive interviewing has become best practice in survey research (Lenzner et al., 2016). Following Goerman's (2018) action points to predict the respondent's difficulties with the questionnaire items, the evaluation focus on the evaluation of indicators, procedures during the survey process, interview recruitment, and the representativeness of the selected countries, given the importance of multicultural studies in this context (Miller and Collette, 2019).

2.1.2.1 Analysis of quality indicators on cognitive testing

An overview of QAP indicators related to cognitive testing show that all quality indicators were successfully met at the CAPI stage.

Table 3. Quality indicators on cognitive testing

Number	Criteria	Indicator	Target	Assessment
64	Coherence & Comparability	The questionnaire and materials tested in the non-source(s) languages have been translated applying TRAPD.	Y	Not applicable Target met for EWCS 2020
65	Accessibility	Evidence of respondents' consent is gathered.	Y	Not applicable Target met for EWCS 2020
66	Accuracy	Percentage of items included in the cognitive test for which systematic documentation is provided about the extent to which answers in the cognitive	100%	Not applicable Target met for EWCS 2020

Number	Criteria	Indicator	Target	Assessment
		interviews correspond with the concepts that are intended to be captured by the question.		
67	Accuracy	Number of questions for which 'major' issues are detected	0%	Not applicable Target met for EWCS 2020
68	Accessibility	Percentage of 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.	100%	Not applicable Target met for EWCS 2020
69	Punctuality	Cognitive test completed in due date	Y	Not applicable Target met for EWCS 2020

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

Indicator 64, related to the Coherence & Comparability, refers to the adherence to the TRAPD translation method which is current best practice to achieve quality translations. This was the case for the Cognitive Test as showed in the Cognitive test report. Indicator 65 related to Accessibility was also fulfilled during the CAPI phase, and evidence for signed consent form from cognitive interviews (consent to participate, consent to audio record and consent to share recording with Eurofound) were gathered.

Accuracy indicators, 66 and 67 were already completed for the CAPI stage, and no new cognitive interviews were required for the EWCTS 2021 questionnaire. Although not cognitive tested, according to the 2021 Pilot Report there were no problems understanding the new COVID-19 vaccination guideline questions, and no new cognitive interviews were needed for the EWCTS 2021 questionnaire.

The second Accessibility indicator evaluates the proportion of questionnaire elements for which there is systematic documentation indicating the degree of alignment between responses obtained during cognitive interviews and the intended concepts to be captured by the questions (as outlined in the glossary). Regarding the Punctuality Indicator (69), the Cognitive test was completed by the due date, and in view of the similarities observed between the two questionnaire versions and the thorough testing that had been previously conducted, there was no need to repeat the testing process as outlined above.

Overall, it can be assessed that the indicators of the cognitive test were successfully met during the CAPI phase, and therefore, the quality of the process maintained. Additional cognitive testing of the EWCTS 2021 reviewed questionnaire, scales, and new questions, together with certain questions that already posed serious challenges across waves even with the use of supporting visual materials would have obviously benefitted the quality of the questionnaire, but Eurofound decision to use already fully completed tasks and adapt them in a very short time was reasonable and well documented given the pressing circumstances.

2.1.2.2 Comparability to ‘gold standards’ and previous EWCS waves

In the field of survey research, the application of robust methodologies and quality criteria is vital to ensure the reliability and validity of the data collected. When it comes to cognitive testing, which involves assessing the comprehensibility and clarity of survey questions, scholars and devoted institutions in Europe have contributed to defining golden standards for achieving a high-quality cognitive testing process (Behr, 2018).

Table 26 of the Appendix presents a comparison of EWCS to gold standards such as the ESS, or Eurostat. Overall, the cognitive test process Eurofound has followed is fairly like top institutions in the field. Even in the aspects in which it differs it still retains the quality. Although improvements could be added in the use of cognitive testing, the ‘3MC’ expert review, and piloting are up to gold standards and have proven critical to the questionnaire development and testing. Several issues were detected in the test with final respondents that could have gone easily unnoticed in full scale pilot. That is different interpretations of the same word, response options that were not considered mutually exclusive by respondents, translation issues, or difficulties in whether or not employing examples (when used participants tend to consider them as a set exclusive list, when not the question was not fully understood).

The implementation of different pretesting strategies from expert review to cognitive testing and piloting, situates the EWCS at the forefront of ‘3MC’ surveys and up to best current standards. However, some issues could be pointed to further improve it.

The comparison of different cognitive test efforts offers some valuable insights. The methodology and form of reporting of the cognitive tests results differs widely across editions. In this vein, the cognitive post test of 2015 is more complete in terms of expertise or experience of the team who carried out the test and in its methodology. Not only was it carried out in three instead of two countries, but it also involved two different techniques including cognitive testing and web probing, and therefore a much bigger sample (365 compared to 40 in 2020). That allows for the test of prevalence of the errors encountered. The report also includes information on the cognitive techniques used, the rationale or objective in the question and the type of probing, with a glossary of techniques with examples. The findings and recommendations are presented per question and refer to sources of error in a TSE framework. It also includes verbatim or direct quotes from participants, which makes the issues detected clearer and less susceptible to interviewer's biases. The cognitive test of 2020 offers, on the other hand, a summary of conclusions and uses a ranking system (ranking an issue from not problematic, somewhat problematic, to very problematic) that makes the prioritisation and study of issues detected more straightforward, it also provides information on which questions were amended or removed after the cognitive test.

The objective of this comparison is not to point at flaws since both exercises were relevant and informed the final questionnaire but to signal good practices that could be carried on, and to point out that differences on both exercises could signal an area susceptible to improvement in the QAP. Despite the methodology used and allowing Eurofound to adapt to the circumstances, both time and budgetary constraints in terms of countries and the corresponding techniques utilised, some standards on how the test should be carried and reported must be established. The reports should include Information justifying the selection of questions and countries.

If the budget allows it, we recommend reinstating web probing. Thus, cognitive interviewing would serve in-depth exploration of new or problematic questions, while web proving can offer some insights on their prevalence and potentially extend the exercise to other countries (Meitinger and Behr, 2016; Scanlon, 2016; Edgar, 2016). Online probing offers some advantages like: Bigger samples, wider variety of respondents, wide geographic scope, elimination of interviewer training, and biases, and elimination of social desirability effects (Behr et al, 2017). It should, however, be noted that some populations, particularly those that already experienced difficulties with the survey, would be more unreachable for both web probing and in a CAWI survey. This is particularly relevant if the next edition is carried as a push to web CAWI, given the unsupervised or lack of assistance in the interview. We strongly recommend to always cognitive test or web probe at least the screening questions, ISCO and NACE classifications, and those that have been repeatedly deemed difficult even if assisted by a person and showcards. That is particularly important when new administration modes are to be applied, whether it is a CATI or CAWI. Focus groups, and debriefing with both respondents or interviewers, could be helpful in finding the right formulation for questions (Campanelli et al, 1991; Morgan, 2005; Gehlbach and Brinkworth, 2011; Haeger, et al., 2012; AAPOR, 2021).

Methodologically, the CAWI allows for the pretesting of different question wordings and cues. It also offers the chance to implement probing in the real data collection in just a few of the questions deemed problematic, which opens the chance to web probe the validity and equivalence of questions in a probabilistic ‘3MC’ sample, and to compare in person cognitive test and web probes across cultures (Behr et al., 2014; Meitinger, 2017).This could, however, have an adverse impact in the accuracy of the data since web probing increases response burden, especially if open-ended, and could potentially affect item nonresponse (on the question and next few questions), survey breaks, shifts in response behaviour, longer answering times, or backtracking (Fowler and Willis 2020; Luebker 2021; Hadler, P., 2023). An alternative would be to apply it to only to a subsample to control these effects (Schuman, 1966; Behr et al., 2017). Closed-ended probes also remain less problematic according to Scanlon (2019) and could be designed from cognitive interviews.

Although cognitive testing can and has detected problems in survey questions, finding suitable solutions is a more difficult endeavour, particularly in ‘3MC’ contexts as changes can create harmonisation and comparability across time issues. This task, and the quest to avoid measurement and equivalence problems, would be further convoluted if mixed methods designs are to be applied in the future. It is important to note that the persistence of issues in certain questions is deemed to be related, not to a poor questionnaire design or insufficient pretesting, but rather with the complex nature of the studied topic. Given the long-standing nature of the EWCS, cognitive testing has well served the EWCS questionnaire. Many of the issues detected in 2015 cognitive tests had been addressed in the EWCTS 2021 questionnaire.

2.1.3 Translation

The EWCTS 2021 questionnaire was translated into 55 languages, including the harmonisation of 4 languages commonly spoken but with different dialects (Dutch, French, German, and Greek) in 9 countries, and the adaptation of source questionnaires from 9 languages in 14 countries which share a similar language or where it is spoken by a minority (Eurofound, 2021a).

The translation processes aimed at ensuring the semantic, conceptual, and normative equivalence between translated versions of the questionnaire. Several measures were taken to this end, including an advance translation performed by two linguistic experts in cross cultural survey translation ('3MC'), external assessment, and a Translation, Review, Adjudication, Pretesting, and Documentation approach (TRAPD), which included cognitive testing in two countries and piloting all 55 languages. The training materials and relevant documentation were also translated using a simplified approach.

Several remarks should be made about this process for the EWCTS 2021 2021. As noted, the EWCTS 2021 translation made use of processes previously carried for the interrupted CAPI phase. The questionnaire developed for EWCS 2020 was translated following a rigorous TRAPD process, in which two translations were produced independently in each language, an adjudicator reviewed them, and then both versions, or a third one if needed, discussed between the adjudicator and the two translators in an online meeting to reach a final agreed upon version, which was additionally reviewed by the fieldwork project manager or team. This translation included the review of previously existing questions to ensure they were up to date with current language, and coherent and consistent with the rest of the questionnaire; the translation of modified questions ensuring their coherence with previous waves, and the full translation of new questions.

Since the EWCTS 2021 used this previously translated questionnaire as basis, with only minor changes having been applied, Eurofound adopted a simplified TRAPD model. Hence, this simplified version of the reworked and modularised source English questionnaire was proofread by the fieldwork contractor, and then one translator translated the new and modified questions and scales, and one adjudicator reviewed this translation. This review took place in writing through the Translation Template, where the adjudicator described the issues encountered. The Translation Template is well-structured, organised, and easy to understand. The comments included in it are relevant, complete, and well-formulated, and the file shows that corrections were done and argued in detail when needed. Discussions, if any, took place via email. There were online meetings for the harmonisation process, albeit only for German and French. Although the decision to apply a simplified TRAPD approach was justified, and in line with quality standards given the circumstances, it is our assessment that a team meeting, even if online, would have been a better approach and more respectful of the team approach to TRAPD translation, which would have ensured the quality to a higher standard at little or no extra cost.

The translation process was, however, thoroughly documented, the questionnaire double-checked by the contractor to ensure the consistency of terms within and across items, and the project managers during the script checking process and pilot. Overall, there were no major issues detected at this stage, and only some issues were encountered particularly in Bosnia and Herzegovina, and Latvia.

The fieldwork materials: the interviewer manual, annotated questionnaire, or glossary, were only translated by one translator and no adjudicator, for efficiency reasons. Other materials were reused from the CAPI phase like the Guidance note on probing, the confidentiality agreement, and data protection notice.

2.1.3.1 Analysis of quality indicators on translation

This section assesses whether the translation process met the accessibility, accuracy, punctuality, and relevance indicators criteria. Because of the complexity of the translation processes, many indicators refer to these tasks. Indicators 60 to 64 refer to the advance translation process, 70 to 73 to the organization, questions to be translated, translators' selection and training, 75 to 77 to the initial translations, 79 to 83 to the within and cross-country adjudication process, and 84 to 85 to the results.

Table 4. Quality indicators on advance translation process

Number	Criteria	Indicator	Target	Assessment
60	Accessibility	Comprehensive documentation of the process of advanced translation	Y	Not applicable Target met for EWCS 2020
61	Accuracy	Percentage of questionnaire items where substantive ambiguities are spotted for which either the source questionnaire is adjusted, or a translation instruction is drafted	100%	Not applicable Target met for EWCS 2020
62	Accessibility	Clear translation instructions and precise documentation schemes for TRAPD are developed	Y	Not applicable Target met for EWCS 2020
63	Punctuality	Advance translation delivered at agreed date to contractor	Y	Not applicable Target met for EWCS 2020

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

All the quality indicators for the advance translation were already met during the CAPI phase, in 2019.

Two experts blindly translated the new or modified questions using a Glossary that listed questions, their rationale, answer options and relevant filters. The questionnaire was reviewed by a survey methodologist with experience in '3MC' surveys. This is, as already stated, an important addition in line with best current standards in the field, which aim is to identify and address translation and cultural issues early on, helping to enhance the translatability of the source questionnaire, and reduce translation problems facilitating its cross-cultural implementation. Although the decision to use the results from the CAPI exercise, given the time and budgetary constraints, is perfectly reasonable, the only drawback is that the expert translators and reviewer could not take into consideration the CATI mode of administration, which might have led to different wording or interviewers' instructions.

In any case, those few indicator items found with potential ambiguities were addressed either in the source questionnaire or by including a translation instruction, based on the guidelines and annotations provided in the "measurement objectives / explanatory notes / translation notes" in the Glossary column and in the Final Transability Assessment. Comprehensive documentation of the process of advanced translation was produced, clear translation instructions and precise documentation schemes for TRAPD were developed produced both by the CAPI and CATI stages, and an advance translation delivered at agreed date to contractor.

Overall, it can be assumed that the process was successfully completed. An important input to endorse this quality process has been the validation by two external experts, who were interviewed and involved in the translation process in the EWCTS 2021. They confirmed that, overall, the effort made by Eurofound to adapt the system from one mode to another has been substantial, aiming to maintain the highest possible quality within the given time frame.

As mentioned, a survey methodologist reviewed the Translation process. Also, one '3MC' expert and the advanced translation team (both members of that team are heavily involved or even led the translation team in the ESS). This expert also confirmed that, although the implementation of advanced translation differs from the ESS, both models are adequate.

Table 5. Quality indicators on translation process

Number	Criteria	Indicator	Target	Assessment
70	Accessibility	Clear process for monitoring the translation and administrating the translation documents is set in place	Y	Target met
71	Accessibility	Percentage of translators and adjudicators whose CVs have been approved by EF in advance.	100%	Target met
72	Punctuality	Decision on questions eligible for translation delivered at agreed date	Y	Target met
73	Accuracy	Percentage of translators and adjudicators that receive training materials prior to commencing work	100%	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

As previously discussed, the Translation Template shows the successful implementation of the simplified TRAPD method, demonstrating effective communication among the involved parties and successful resolution of any issues (Eurofound, 2021b).

Most of the translators were already involved in the EWCS CAPI phase, and Ipsos had only to provide 5 new CVs for EWCTS 2021. We could assess a template that shows that all CVs of translators and adjudicators were approved by Eurofound, complying with the requirement of being native speakers of the target language, fluent in English at a C2 level, and with experience of translating questionnaires and other materials for market or social research, or previous experience in the EWCS. It also shows that issues with translations were detected and addressed, for example by appointing a third translator in Slovakia. This log is for the 2020 process, and we could not find information updated on the five new translators for 2021. The approval of CVs by Eurofound was, however, also confirmed by interviews performed by the quality assessment team.

The deadlines were successfully met. One of the interviewed experts claimed that the process, as well as the meetings necessary for validating modifications or addressing other issues in the process, were satisfactory, and highly efficient. Although there were minor delays in deliveries, these did not impact the overall translation process.

All translators received a pack of concise briefing documents prior to the translation process, with the translation log, background information on the survey, a brief user guide for translation, adjudication and, where applicable, guidance for the harmonisation or adaptation, a document with additional checks to be performed and the glossary produced in the advance translation. Notwithstanding there is no indication of the glossary being delayed for a couple of days. In this regard, the indicator is somewhat vague as it does not state which documents are compulsory or constitute the training materials. Nevertheless, the translators had sufficient information to provide an adequate translation and the quality was ensured.

Table 6. Quality indicators on initial translation

Number	Criteria	Indicator	Target	Assessment
75	Accuracy	Percentage of countries where translation is carried out by a professional translator	100%	Target met
76	Accessibility	Percentage of countries for which systematic documentation of results of initial translation (in accordance with template) is provided	100%	Target met
77	Punctuality	Initial translation delivered at agreed date	Y	Target met

Source: Author’s own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

In line with indicator 71 all translations were developed by a professional translator, and therefore we consider indicator 75 as fulfilled.

The results of the initial translations and the adjudication were systematically documented and can be easily traced in the documentation and translation log. The Excel translation file showed that the simplified TRAPD method had been applied correctly and the translators involved had, where required, argued their cases in sufficient detail. This meticulous approach ensured transparency and clarity in the translation process, with all translation files containing organized documentation. The Translation Report (Eurofound, 2021b) shows all countries along with their corresponding translators/adjudicators who have conducted the translation process in each country. Language Connect, Ipsos' translation partner, took on the role of overseeing the translation process at a local level. Whenever challenges arose or there were questions, the project manager from Language Connect communicated with Ipsos CCT for resolution. In conclusion, the quality of the indicators assessed in the Quality Assurance and Control Report, conducted by Ipsos in 2021, demonstrates a high level of diligence and systematic documentation.

Table 7. Quality indicators on adjudication

Number	Criteria	Indicator	Target	Assessment
79	Accessibility	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
80	Punctuality	Within country adjudication (overall) delivered at agreed date	Y	Target met

81	Accuracy	Percentage of cross-national review sessions (possible via email), in which adjudicators from each of the countries sharing the particular language participate	100%	Target met
82	Accessibility	Percentage of countries for which systematic documentation in English is provided about the process and results of the cross-national review (in accordance with template)	100%	Target met
83	Punctuality	Cross country review (overall) delivered at agreed date	Y	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

As stated above, the adjudication process both, within and across countries, can be easily traced in the Translation Template and documented in the Translation Report. The adjudicator provided a concise explanation in English about the decisions made during the review discussions. Although the indicators have been fulfilled, we consider that the accuracy indicator 81 is vital to ensure the quality in the translation and is too lax to be considered in line with the team-oriented approach of the TRAPD even in its simplified form. We therefore strongly suggest reviewing it to ensure the Review step is up to best current standards.

Table 8. Quality indicators on final translation

Number	Criteria	Indicator	Target	Assessment
84	Accuracy	Percentage of questionnaire items that required editing (e.g. correcting typo's, copying and pasting errors, etc.)	0%	Target met
85	Punctuality	Final translated questionnaires (language version) delivered at agreed date	Y	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

All editing of items in the translations were carried out before the main fieldwork started. In some minor occasions some reviews went through only one translator, and the fieldwork project manager, such as in the case of the reviewed introductory text. There is no indication on a delay of the final translated questionnaires.

2.1.3.2 Comparability to 'gold standards' and previous EWCS waves

The translation process for EWCTS 2021 has consistently met quality standards, aligning well with "gold standards" in the field. Furthermore, it can be noted that Eurofound has increased their efforts to maintain the quality and equivalence of questions updating their methodology and standards to meet best current practices in '3MC' surveys and previous assessments recommendations.

An example of this is the incorporation of an Advance Translation process before the CAPI phase which aims to identify and address translation and cultural issues early on, improving the translatability of the source questionnaire and facilitating its cross-cultural implementation. Creating high-quality survey instruments involves a rigorous process that ensures accuracy, comprehension, and cultural sensitivity of the questions. Advance translation is a critical step in this process, which aims to translate survey materials accurately and effectively into multiple languages while maintaining the original intent and meaning. Eurofound conducted the original advance translation in 2019 before the CAPI fieldwork. Two experts independently translated the new or modified questions using a glossary provided by Eurofound, which included question details, answer options, filters, and other relevant information. Translators also had the opportunity to comment on their translations. One of the focuses was on assessing syntactic and grammatical suitability for translation. Also, the main point was identifying items which would be hard to translate because of cultural and language differences around European countries. In addition, a survey methodologist with experience in '3MC' surveys reviewed the questionnaire, particularly the newly constructed questions and items. The advanced translation contributed to developing specific instructions, training materials for translators, and to inform about the inputs during source revision. The review process resulted in changes to question-wording, drafting interviewers' instructions, clarifying certain topics, reordering questions, and identifying potentially ambiguous wording. The questionnaire and translation files provided clear instructions, item-level notes, definitions of complex concepts, and the objectives of each question.

As thoroughly discussed, the EWCTS 2021 questionnaire followed a simplified TRAPD approach since it presented only minor changes from the previously fully translated, reviewed, adjudicated, pretested, and documented questionnaire for EWCS 2020. Although the decision to apply a simplified approach was justified, and in line with quality standards given the circumstances, and despite the fact that it could be argued that the questionnaire was in fact translated by three different translators, it is our assessment that the team approach of the Review and Adjudication processes in the TRAPD was not fully endorsed, and that a team meeting even if online would have ensured the quality of the translation and harmonisation to a higher standard at little or no extra cost.

The translation process for EWCTS 2021 201 maintained a high quality, despite circumstances. Eurofound's plans aligned with quality standards, showing a proactive approach. The process carefully ensured inclusivity, clear sentence structure, and preserved complex concepts across 36 nations and 55 languages, highlighting a commitment to translation excellence in challenging circumstances. The resulting questionnaire was pretested both at a cognitive test (for the 2020 questionnaire) in a full-scale pilot conducted in all countries and the whole processes thoroughly documented.

2.2 Fieldwork

As already stated, the CAPI fieldwork for the EWCS had to come to an end after seven weeks due to the spread of COVID-19. This assessment primarily focuses on the subsequent 2021 CATI fieldwork, aiming to evaluate the entire process and identify areas for improvement in future rounds. It also highlights sections less affected by the transition from CAPI to CATI but relevant for potential CAWI implementation. The approach prioritizes addressing issues affecting results, highlighting successful aspects, and providing improvement recommendations for future survey rounds.

2.2.1 Analysis of quality indicators on the fieldwork process

The pre-fieldwork pilot process is key for identifying and reducing measurement errors that can harm population-level statistical estimates. This process is a condition for maintaining comparability across populations in '3MC' surveys (Survey Research Center, 2016). Pretesting entails various activities to evaluate the effectiveness of survey instruments, data selection and collection, and overall field procedures.

The analysis of indicators in the '3MC' literature emphasises special considerations. Hambleton, Yu, and Slater (1999) highlight several factors to account for in diverse cultural contexts during survey application. These include evaluating interview length, adapting the instrument, assessing population, familiarity with measurement units, testing comprehension of constructs and concepts, reviewing instrument design, and understanding response processes and behaviours.

When multiple languages are used in a survey, pretesting different language versions becomes important to ensure measurement and culture (Devins et al., 1997) and cross-cultural equivalence (Hui and Triandis, 1985). Additionally, employing the same data collection mode across countries in a cross-national project can be challenging. Pilot techniques may have limitations in specific contexts and cultures. Hence, it is compulsory to test the adaptation of strategies systematically and interactively in diverse populations. Collaboration between survey implementers and commissioning entities is vital (Pennell et al, 2017).

Table 9. Quality indicators on fieldwork (pilot)

Number	Criteria	Indicator	Target	Assessment
86	Accuracy	Percentage of countries where pilot interviews are carried out with at least 40 respondents	100%	Target mostly met
87	Accessibility	Percentage of countries where pilot interviews are carried out in all local languages	100%	Target mostly met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

The first indicator (86) was achieved in most countries. There were a few countries that exceeded the target by conducting more than 40 interviews (e.g., Romania completed 62 interviews). The only country that failed to achieve the minimum number of interviews required was Spain, as it conducted a total of 39 interviews. This number is nearly identical to the minimum number of interviews required per country, so it can be said that the accuracy indicator and more importantly the assurance of quality was adequately fulfilled. Regarding the second indicator (87), Ipsos Spain faced challenges in conducting interviews in Catalan despite their efforts during the allocated pilot fieldwork period. Since the survey follows a random probability approach, it was not possible to establish language quotas.

In some cases, refusals were due to these language barriers, such as in the Czech Republic, Germany, Finland, Cyprus, and Luxembourg, where the fieldwork contractor signalled a loss of 10% of the interviews due to language barriers. Including these populations, especially in countries with significant immigrant populations, is relevant to reduce measurement error and nonresponse but also to address vulnerable groups likely more affected by worse working conditions. It should be noted that the EWCS already increased notably the number of languages from previous editions. But the same approach used in the CATI for Luxembourg and Cyprus where an English version was added to reach people working in the country at EU institutions, could be applied in other countries, particularly for already translated languages. CAWI in this sense offers the possibility to answer the questionnaire in any of the languages scripted regardless of the country, for example 14,5% of the population of Luxembourg which could answer in Portuguese, or many workers from non-European countries but fluent in one of the languages of Europe because of colonial legacies.

Table 10. Further quality indicators on fieldwork (pilot)

Number	Criteria	Indicator	Target	Assessment
88	Accuracy	Percentage of scripting errors detected in the pilot test for which a solution is implemented and tested	100%	Target met
89	Accuracy	Percentage of countries where pilot interviews cover the six questionnaire modules	100%	Target met
90	Accessibility	Pilot test completed at agreed date	Y	Target not met
91	Accessibility	Comprehensive pilot report provided	Y	Target met
92	Accuracy	Percentage of items in the source questionnaire changed after pilot	0%	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

The target for indicator 88, which pertains to resolving issues in the pilot script, was successfully met. Detailed solutions were agreed upon for each minor issue encountered, and these are described comprehensively in the Ipsos Pilot Report (2021b).

The second indicator (89) was fulfilled incorporating random modules of completion, which were automatically distributed by the computer system. This allowed for proportional distribution of completed surveys across the modules. However, certain issues were observed when interviewers failed to close surveys properly, causing the system to not recognize that a module's quota had been fulfilled. As mentioned in the pilot reports Ipsos (2021b), had to conduct a secondary check of completed quotas for each module at the end of daily fieldwork to ensure random completion of the modules. Although Ipsos solution to this problem was effective and clever, it involved duplicating efforts to achieve complete questionnaires with similar module representation.

In a post-hoc evaluation, a filtering system for the modules could have been a better alternative, but implementing such a system would have required modifying the questionnaire. Therefore, we consider the adopted solution to be efficient and accurate. However, it is important to acknowledge that this may affect the representativeness and comparability with previous editions of the survey, besides the already mentioned limitations considering the CAPI-CATI implications in this regard. Consequently, in terms of capturing information from all modules of the questionnaire, the CAPI system collects significantly more data per respondent, at the expense of longer completion times.

The pilot was originally scheduled from November 30 to December 20, 2020, but minor delays occurred in some countries due to the late reception of translated scripts. The last country to start the pilot was Poland on December 4. Overall, the pilot lasted an average of 18 days, ranging from 9 to 30 days depending on response rates. Twenty-one countries completed the pilot by December 20, while fifteen countries finished on December 31, not fully meeting indicator 90. Additionally, Bosnia Herzegovina and Sweden mistakenly recorded minimum call attempts after January 4, 2021, concluding fieldwork on December 31, 2020, due to an error. However, this did not significantly impact the survey's quality.

Eurofound received a preliminary version of the pilot report on January 25, which provided enough information to make necessary adjustments to the main fieldwork. An online meeting was held between Eurofound and Ipsos to discuss the conducted pilots and finalize plans for the actual fieldwork. Given the prioritization of preparing for the main scenario, the final version of the report was submitted on July 9. Eurofound monitoring teams consider indicator 91 to be met.

Upon analysing the Technical Report (Ipsos, 2021c, p. 102-103), certain issues regarding the duration of the Pilot Test were identified. Firstly, the response rate in some countries was lower than expected. These rates should be considered for future implementations of a CATI survey. According to Ipsos technical report, technical problems (as the guidance note on probing be made more concise as it is currently too detailed, the interviewer and the trainer manual) have some consequences in the low response rate in Finland and Switzerland. As for Sweden, it is attributed to the typical low response rates observed in Scandinavian countries (Christensen et al., 2022).

Adapting a CAPI questionnaire to CATI is always a challenging task. While the adaptation work was well-executed, it is evident that the questionnaire has undergone modifications, which can potentially pose difficulties when comparing data with previous survey waves. However, if the core indicators of the questionnaire have been maintained, as they have, the survey remains valuable in terms of its primary objective to compare the employment situation across countries in 2021 amidst the significant changes brought about by the pandemic.

That said, a twenty-minute duration for a CATI survey is relatively long, requiring a well-trained team of interviewers. Thus, we consider the preparation and training of interviewers to be relevant for the successful implementation of this type of survey. We also find the recommendations in the Ipsos technical report on piloting to be valuable. Local agencies suggest emphasizing confirming respondent eligibility as much as possible during the field team training process to minimise misunderstandings during the final survey implementation.

2.2.2 Fieldwork infrastructure

The fieldwork infrastructure encompasses the processes involved in survey implementation. There are two issues to consider in a ‘3MC’ environment: on the one hand, the laws in different countries regarding automated call systems; and in second place, the disadvantages, and advantages of using the different CAPI, CATI and CAWI information collection systems in countries with co-official languages and with strong sociocultural differences (Survey Research Center, 2016). Most indicators used in this phase are specific to the CATI administration. As a result, establishing comparability with previous survey applications becomes challenging. It is important to note that these indicators may be subject to modification in future transitions to alternative methodologies such as CAWI.

However, taking into account the homogeneity of the 36 countries in terms of their telephone infrastructures and administrative regulations, as no country in Europe has any restrictive regulations, such as the United States (Survey Research Center, 2016), which does set certain restrictions on the use of different automated calling systems, it is possible in the EWCTS 2021 to use all kinds of tools to optimise a common platform for the CATI data collection system.

Table 11. Quality indicators in fieldwork infrastructure

Number	Criteria	Indicator	Target	Assessment
93	Coherence and comparability	Percentage of countries using an automated dialling system for contact management	100%	Target met
94	Coherence and comparability	Percentage of countries using a common platform for collection interview data	100%	Target met
95	Accuracy	Scripting is tested and hard and soft data checks are integrated	Y	Target met
97	Accuracy	Percentage of countries using CATI web links ('CATI Links') for survey data collection	25%	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

Eurofound utilises a total of four indicators to assess this infrastructure. Coherence and Comparability system indicators (93 and 94) measure the percentage of countries utilising an automated dialling system for contact management and a common platform for collecting data. The first Accuracy indicator (95) assesses whether the survey script was reviewed by agencies and if checks for both hard and soft data integration were conducted. While the second Accuracy indicator (97) measures the percentage of countries utilising CATI web links for survey data collection. Both Accuracy Indicators evaluate the accuracy dimension of the fieldwork infrastructure.

To ensure accuracy and consistency in the fieldwork infrastructure, Ipsos implemented a data collection system called Dimensions, which has been the preferred survey scripting and data processing platform globally for several years. This integrated platform offers centralised functionality for data collection, scripting, and sample management across multiple countries. By using this system in all countries, Ipsos achieved consistency between the dialled phone numbers and the collected data, fulfilling the requirements of Coherence and Comparability Systems indicators.

The script underwent a thorough review by Ipsos and Eurofound, and daily automatic validations were set up. Soft and hard checks were implemented to meet the Accuracy requirements of the script. Once the questionnaire was revised and transformed from CAPI to CATI, the script was configured. Ipsos generated a single script that was used across all 36 countries with appropriate translations.

The script's data layout was designed based on the field tool structure to minimize changes during the transition from CAPI to CATI on different platforms. The scripting team meticulously reviewed question wording, filters, randomisations, value restrictions, and more. They conducted extensive testing to ensure filters and question paths worked correctly, answer options were displayed accurately, and instructions were clear. Ipsos also performed tests with dummy interviews to validate complex question routes and modulation behaviour.

The second Accuracy Indicator was met but there was a differentiation between countries either accessing the survey via web links or directly. (Due to the variety of local systems involved in the project, Ipsos found that there is wide variety in the outcome codes used or available for local teams, while having the same reporting meaning). The CATI Links platform was used by monolingual countries, while the CATI Direct option was used by multilingual countries. For multilingual CATI Direct countries, the platform allowed language selection at the beginning of the interview. This meant that the contact procedure could be executed in two ways:

- Contact procedure with bilingual interviewers: During the initial contact, the respondent's preferred language was identified, and the interviewer marked it in the platform. The changes were loaded with translations for all surveys, allowing the interviewer to conduct the interview in the selected language.
- Contact procedure without bilingual interviewers or interviewers with low proficiency in other survey languages: During the initial contact, the interviewer marked the interview for recontacts and changed the default survey language to the next survey language, or the interview was marked for recall after changing the language preference in the survey. In both cases, the interview would be conducted by a different interviewer proficient in the selected survey language.

2.2.3 Data entry

The data entry process largely depends on the software and data collection system used. For this survey, a single company managed the fieldwork implementation, and data verification was carried out using a uniform system.

This worked in favour of the quality contributing to a uniform and compatible fieldwork development, script integration and data validation. As discussed in the previous section all countries used Dimensions software, unlike in the previous edition where different systems were used, although some used their local interview systems and accessed it via web links. Dimensions was also used for sample management and the collection of paradata.

Some issues were detected however regarding the variety of outcome codes and data delivery by local partners. But overall, the use of a single company and system ensured a higher level of accuracy and coherence and compatibility. The Precision and Accessibility indicators evaluate the validation of the number of hard consistency rules programmed in CATI and the number of soft rules. The remaining Accessibility indicators assess the understanding of available documentation on both hard and soft rules during the interview and data verification stages.

The analysis aimed at evaluating the consistency of the data entry verification test.

Table 12. Quality indicators in data entry

Number	Criteria	Indicator	Target	Assessment
99	**Accuracy**	Number of hard consistency rules identified	**>0%**	**Target met**
100	**Accessibility**	Comprehensive documentation of all hard consistency rules	**Y**	**Target met**
101	**Accuracy**	Number of hard consistency rules identified and programmed in CATI	**>0%**	**Target met**
102	**Accuracy**	Number of soft consistency rules identified	**>0%**	**Target met**

Number	Criteria	Indicator	Target	Assessment
103	Accessibility	Comprehensive documentation of all hard consistency rules	Y	Target met
104	Accessibility	Comprehensive documentation of all soft consistency rules	Y	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

Ipsos prepared and identified 72 hard consistency rules for the verification of the data. These rules were implemented during the final SPSS data validation procedures. The checks performed ensured that the data was collected correctly with no additional or missing data collected. For the pilot and in the main stage, a total of 23 soft controls were identified and agreed with Eurofound. These controls were implemented directly into the CATI scripts with interviewers receiving notifications when inconsistencies were identified.

Hard rules were set up through filters that allowed the automatic routing of surveys and blocks. For the age variable, a series of controls were set up so that respondents under 16 or outside the correct age brackets were not surveyed, nor were respondents who were not working automatically selected.

2.2.4 Interviewer training

When conducting '3MC' surveys, the significance of interviewers should be considered. Undertaking a pilot process provides project leaders with the opportunity to assess the performance of interviewers, which offers an advantage. In cases where interviewers are unsure about the procedure, retraining can be implemented. In certain circumstances, it is even advisable to train an excess number of interviewers. Consequently, we emphasize the importance of considering these factors for future applications of the survey, particularly in contexts where there is a CAPI or mixed application (Lyberg, et al., 2021).

The training and documentation of interviewers is essential in the survey processes, especially in the '3MC' framework. Interviewers are often required to multitask with a high level of precision related to a series of skills that need to be trained. Among others, we could cite the selection of respondents, motivation, communication, obtaining adequate answers for their treatment (especially in this type of language and cultural immersion surveys). It is especially important as far as possible to avoid the influence of the interviewers, -"interviewer effect"- in obtaining responses. All efforts in the training and documentation of the interviewers are aimed at minimising the effect that their behaviour could have on the respondents which could lead to sampling error, non-response error, measurement error and processing error through inadequate training.

It should be noted that these phases of '3MC' surveys play a relevant role in obtaining comparable data. The literature identifies this phase of survey implementation as one of the most challenging (Lyberg, et al., 2021). The extreme cultural and contextual variability between countries is even greater in face-to-face surveys. The training phase is therefore important to reduce the risks associated with possible interviewer-generated rejection. Indicator 105 relates to the attendance of the national fieldwork managers/representatives at the webinar briefing meetings that were held prior to the pilot and mainstage fieldwork.

Table 13. Quality indicators in interviewer training

Number	Criteria	Indicator	Target	Assessment
105	Accuracy	Percentage of national fieldwork managers or representatives attending the pre-pilot and mainstage fieldwork webinars	100%	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

According to Ipsos technical report, the training sessions (TtT) were successfully conducted as scheduled, and no issues were recorded during these sessions. The feedback collected by Ipsos regarding the training sessions and the materials, including PowerPoint presentations, was predominantly positive, indicating that participants rated them favourably.

Alongside the positive feedback, Ipsos also gathered suggestions for improvement that can be implemented in future editions of the training sessions. These valuable points for improvement will help enhance the training experience and ensure continual refinement of the materials and delivery methods for subsequent sessions.

2.2.5 Training materials

When it comes to materials, careful consideration should be given to determining the most precise and suitable options based on the type of survey application. It is essential to assess which materials offer the broadest coverage across multiple countries and to explore ways to enhance the instructions for improved clarity regarding the questionnaire's contents.

Table 14. Further quality indicators in interviewer training

Number	Criteria	Indicator	Target	Assessment
108	Accuracy	Training materials cover strategies for convincing reluctant respondents	Y	Target met
109	Accuracy	Training materials cover guidelines on contacting process	Y	Target met
110	Accuracy	Training materials cover instructions on CATI program/questionnaire	Y	Target met
111	Comparability	Training materials cover international classifications (ISCO, NACE and ISCED) and implications in terms of respondents probing	Y	Target met
112	Accuracy	Training materials cover the content of the questionnaire	Y	Target met
113	Accuracy	Training materials cover instruction on consistency checks	Y	Target met
114	Accessibility	Percentage of countries for which all training materials are provided	100%	Target met
116	Accuracy	Training covers all relevant materials	Y	Target met

Number	Criteria	Indicator	Target	Assessment
117	Accuracy	Percentage of interviewers that take part in the training	100%	Target met

Source: Author’s own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

The indicators (108-117) assess the documentation of the training. The indicators grouped in the Accuracy dimension assess whether the documentation covers the aspects related to contacting respondents, the guide to contacting respondents, the CATI program instructions, the content of the questionnaire, and the instructions of the consistency checks. They also measure whether the training covered the relevant materials and the percentage of interviewers who took part in the training process. All indicators have been successfully met.

2.2.6 Interviewer selection and briefings

All agencies involved in the study used experienced interviewers who were acquainted with survey research. Importantly, all interviewers had excellent language skills and were native speakers of the language(s) of their respective countries.

Each assigned interviewer received full training before starting the fieldwork. This training was provided by the project or field managers and lasted a minimum of two hours. To ensure safety, especially given the pandemic situation and the need for social distancing, most agencies provided training to their staff through webinars. These measures demonstrate the agencies' commitment to ensuring that interviewers were properly trained and prepared to conduct the survey effectively and accurately.

Ipsos Technical Report contains comprehensive information about the data, providing detailed insights. As for the materials used in the training sessions, the following resources were utilized: a Power Point interviewer training manual, an annotated questionnaire, a guide to assist interviewers in understanding the importance of collecting detailed information, a Data Protection Notice, and coding instructions. These materials were employed to support and guide interviewers throughout the survey process.

2.2.7 Fieldwork phase

During the fieldwork phase, two most important elements come into play: the interaction between interviewers and respondents, and the supervision conducted by either the fieldwork company, external controllers such as Eurofound, or both. This supervision can occur in real-time or through subsequent checks of the fieldwork.

In the contact phase, a key point is to evaluate how the final sample is being obtained and if the target number of interviews are being fulfilled. Additionally, the adherence to fieldwork rules, the number of unsuccessful and validated contacts, as well as the identification of favourable days and hours, when the most interviews where achieved, are assessed. In terms of supervision, it is essential to control the fieldwork teams, monitor the number of interviews conducted by each interviewer, address any issues that arise, find solutions, review progress, and ensure the objectives are being met.

In this process it is important to take into consideration future implementations of technological advances, especially in CAPI methodology (Lyberg, et al., 2021). The availability of cost-effective devices and user-friendly software, which can facilitate comparable quality control processes across study countries, is enabling a real revolution in approaches to quality control in face-to-face '3MC' surveys (Seligson and Moreno Morales, 2018).

Table 15. Quality indicators in contact phase (fieldwork)

Number	Criteria	Indicator	Target	Assessment
119	Accuracy	Percentage of gross sample entries that are discarded before the net sample is realised, for which a final outcome has been realised	100%	Target not met
120	Accuracy	Percentage of valid sample entries that were contacted according to fieldwork rules	100%	Target not met
121	Accuracy	Percentage of interviewers carrying out less than or equal to 200 interviews	100%	Target met
122	Accuracy	Percentage of interviewers carrying out less than or equal to 200 interviews	100%	Target mostly met
122	Accuracy	Percentage of issues identified based on information in weekly monitoring data for which a solution or explanation is provided by the contractor	100%	Target met
123	Accuracy	Percentage of countries where 10% of all contact attempts and 10% interviews are checked	Y	Target met
125	Accessibility	Percentage of countries covered in weekly monitoring data	100%	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

In this section, the first two indicators focus on measuring the contact phase, specifically in relation to the Accuracy dimension. The first one quantifies the percentage of initial sample entries that are discarded before the net sample is determined and an outcome is achieved. On the other hand, the second one indicates the percentage of valid sample entries that were successfully contacted, adhering to the specified fieldwork rules.

The other indicators, assess fieldwork monitoring. These indicators measure the percentage of interviewers conducting less than 200 interviews, the percentage of issues identified on a weekly basis for which a solution has been identified, and the percentage of countries where 10% of contact attempts and 10% of interviews were checked. While these indicators measure aspects related to Accuracy. The remaining indicator measures Accessibility. In this sense, this indicator measures the percentage of countries covered in weekly monitoring data.

Ipsos implemented an interview duration calculation process specifically for this project, and the corresponding documentation is detailed in Ipsos Technical Report (2021c). This process consisted of capturing duration in specific questions, excluding certain elements from the total count.

In particular, the following elements were excluded from the duration calculation: the survey introduction screen, the question on the respondent's age, questions related to information on the respondent's behaviour and the questionnaire review. These exclusions were applied to obtain a more accurate estimate of the actual duration of the interviews by focusing on the questions and answers directly relevant to the study in question. The documentation in the Ipsos Technical Report provides more details on how this calculation was carried out and the specific considerations considered.

Ipsos conducted an analysis to pinpoint questions affecting interview length. Respondents agreeing to be contacted again (Q13) extended the interviews, and follow-up questions were time-consuming, averaging 153 seconds with a median of 138 seconds. The final survey question also took considerable time, averaging 93 seconds with a median of 46 seconds. In 14 cases, interviews lasted 45 minutes or more due to these questions. This underscores the need for effective time management during data collection, considering the impact of these questions on interview duration.

Ipsos identified that another reason for long interviews was the low engagement of respondents, which led to the need for additional clarifications during the interview. In addition, it was observed that in older respondents, these factors may increase.

The contractor addressed excessively short interviews in various countries. Slovenia took measures like withdrawing two interviewers and adding 436 interviews due to some short interviews. In Cyprus, 28 short interviews were found, but no specific measures are mentioned. Greece identified 9 such interviews, with detailed reasons provided, and conducted a thorough review, removing erroneous interviews from the database.

These measures reflect Ipsos' efforts to address abnormally short interviews, either by taking specific actions, such as replacing interviewers in the case of Slovenia, or by conducting thorough reviews and removing erroneous interviews in other cases, such as in Greece. This demonstrates a commitment to ensuring the quality and reliability of the data collected during the survey.

Partial non-compliance with the indicator related to the set limit of 200 interviews per interviewer has been observed in some countries due to the need to re-interview for reasons such as correcting modularisation or replacing invalid interviews. Italy, Montenegro, and Slovenia have had more interviewers exceeding the limit, with a total of 4, 3 and 3 interviewers respectively. In the rest of the countries, the number of interviewers exceeding the limit does not exceed two. This partial non-compliance occurred during the main fieldwork, while in the pilot process the limit was met.

During the fieldwork, a total of six documents were used as documentation. These documents include the Data Protection Notice, the Interviewer Manual (which includes the training slides used during the training phase), the Interviewer Confidentiality Agreement, the Interviewer Training Attendance sheet, and the Interviewer Feedback Form.

The feedback collected revealed similar appreciations to those received during the piloting and training phase. For example, local agencies in Croatia and Finland suggested that the survey question guide could be more concise, as could the interviewer's manual. On the other hand, the French and Luxembourg agencies pointed out that the training manual is more useful for supervisors than for interviewers due to its excessive length. They suggest that the manual could be more efficient if the number of pages were reduced. In addition, Portugal recommended including more instructions on open-ended questions.

2.3 Sampling

This chapter outlines the sampling approach for EWCTS 2021. The framework adopted in this work for assuring and assessing quality is the TSE (Groves, et al., 2011). The TSE paradigm is widely accepted as a conceptual framework for evaluating survey data quality and TSE is a valuable framework for comparative studies (AAPOR, 2021). The TSE framework helps organize and identify error sources and estimates their relative magnitude, which can assist those planning '3MC' surveys (Johnson, et. Al, 2018) to evaluate design and implementation trade-offs.

In this section the focus is on the measure of representativeness of the sample, i.e., whether one can generalize to the target population using sample survey data. For this, the following must be considered: coverage error, sampling error, non-response error, and adjustment error.

Based on the analysis of Quality Assurance and Control Report (Ipsos, 2021a), Technical Report Ipsos (2021c), Sampling and Weighting reports (2021d) for the EWCTS 2021, and all the documentation before and during the fieldwork provided by Eurofound, a series of comments are presented based on the discussion of the results of the indicators and the in-depth analyses of their evidence. Technical reports from previous surveys have also been considered, although in this edition of the survey, the selection of the sample in each country and the reweighting of the sample have been carried out using different procedures than in the previous edition. Accordingly, recommendations considered relevant for upcoming EWCS surveys are provided and further developed in the last part of the report.

2.3.1 Analysis of quality indicators on sampling

The EWCTS 2021 aims to be representative of the population of all individuals aged 16 and over, whose usual place of residence was in the territory of the country and who did at least one hour of work for pay, profit or family gain, for money or other payment in kind in the last week. The survey covered 36 countries, where random probability designs were used to draw samples from the population of each country, although the procedures were not the same across all countries.

The sampling strategy is properly documented in the Sampling and Weighting report and the Technical Report; therefore, the information would not be repeated and only some key elements of the sampling strategy would be mentioned.

2.3.2 Sampling

All participating countries implemented a probability-based sample design, using a high-quality sampling frame, and developed sampling strategies with the objective of minimising sampling errors and therefore maximising efficiency. It must be noted that two largely different sampling designs were used: in Sweden, a stratified sample by LAU2, gender and age with proportional allocation was selected from the sampling frame. In all of the remaining countries, however, a simple random sample was selected by Random Digit Dialling (RDD). It must be noted that this simple random sampling procedure is not widely used in relevant '3MC' surveys; this can be seen as an advantage given that many more complex sampling designs, that might be more convenient in some situations, usually contribute to inflate the variance of the estimates.

Eurofound's QAP includes several accuracy and punctuality indicators to assess elements relating to the sampling frame, the sampling plan, and the sample size.

2.3.3 Sampling frame development

Sampling frames were obtained differently for these two groups of countries:

- In Sweden, the population register was used as the sampling frame for the EWCTS 2021.
- In the remaining countries, RDD was used as the sampling frame. The RDD samples were generated using the most recent lists of mobile prefixes allocated in each country. The survey coverage depends thus on the level of mobile phone use.

All targets were fully met except for the first one, regarding the coverage of the sampling frames.

Table 16. Indicators in sampling frame development

Number	Criteria	Indicator	Target	Assessment
3	Accuracy	Percentage of countries where the sampling frame covers at least 95% of the populations	100%	Target not met
4	Accuracy	Sweden only (register sample): Percentage of sampling frame units for which the contact information was incomplete, and which were not contacted using other means.	2%	Target met
5	Accuracy	In countries using a register sampling frame (Sweden), percentage of sampling frame units that refer to non-eligible addresses.	6%	Target met
9	Accessibility	Percentage of countries for which the characteristics of the sampling frame and procedure are documented in complete accordance with the template (based on Terms of Reference).	100%	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

The first three indicators (indicators 3, 4 and 5) are related to the quality dimension of Accuracy that is particularly critical to sampling, as it reflects whether the sampling frame is an accurate reflection of the population. However, only the first indicator refers to all the countries that participate in the survey, while indicators 4 and 5 mention a single country. The last indicator (number 6) is related to Accessibility.

Indicator 3 is the most important one as it focuses on the sampling frame coverage of all countries. Coverage bias is one of the most important non-sampling errors that can impact the survey representativity. Quality sampling frames for mobile phone-only surveys were sought in each country: all RDD sample frames were able to provide full coverage of mobile phone users in each country, and hence the survey coverage depends on the level of mobile phone use for the target population (working population aged 16 and over). Eurobarometer data on mobile telephone use among the employed population indicates that, in the vast majority of countries, the proportion of individuals that are not covered by the sampling frame is less or equal to 1%. According to Ipsos, 2021d EWCTS

2021 Sampling and Weighting report, in some countries this percentage can be up to 5% (the case of Romania), and other countries do not have available data. In these cases, other official statistics sources have been used, but it must be noted that no official information could be found for Switzerland. On the other hand, coverage errors can be caused by other factors (people not covered by the sampling frame, or people who asked to be excluded from telephone research studies). According to the calculations of the contractor, the target of having at least 95% of the population of interest covered was only achieved by two countries, which means that there is a significant difference between the target value and the actual value in this important indicator. In Sweden, where a different framework was used, coverage targets were not met either, being 84%.

The estimated coverage in some countries is below 90%, which could result in important coverage biases since the use of mobile phones could vary considerably by age, working conditions and other sociodemographic characteristics, something that has been pointed out by various studies (Blumberg and Luke, 2020; Pasadas-del-Amo, 2018). Although data on the mobile phone coverage for the sampling units of this study (individuals) was not available by demographic strata, some official statistics can be found for mobile phone equipment in households. For instance, the percentage of households with a mobile phone in Spain is 99.5%; however, this percentage drops to 98.7% for households with an income lower than 900 EUR, while raises to 99.9% for households with an income higher than 3,000 EUR (INE, 2022). Regarding the effect of age, data from Germany reveals that 100% of households whose main income earner is between 18 and 35 years old have a mobile phone, while this percentage drops to 98.5% for earners between 55 and 65 years old, 98.2% for earner between 65 and 70 years old, 96.5% for earners between 70 and 80 years old, and 89.4% for earners older than 80 years (Destatis, 2022). It would be recommendable to study the key characteristics (sociodemographic and working conditions) of the uncovered population in the countries with the higher noncoverage rates. However, it should also be noted that the potential biases caused by undercoverage in this survey might be relatively small, given that the size of the differences in coverage across demographic groups in the official statistics consulted are, in many cases, of a few percentage points.

Changes in population coverage patterns are the reason why dual frame sample surveys are becoming more common in survey practice. Dual frame surveys (Lohr, 2009, Lohr and Rao, 2006) that combine RDD telephone samples and face-to-face samples emerged to reduce cost, and noncoverage, and could be an acceptable solution for countries with large noncoverage rates. Estimation is not straightforward in dual frame surveys due to the overlap between the two frames. Since their introduction (Hartley, 1962), dual frame surveys have gained much attention and several estimators have been formulated based on several different approaches (Mecatti, 2007). Calibration for dual frame surveys has been studied in Ranalli et al. (2016) and Molina et al. (2015), proving that the bias of the calibration estimates is negligible and reduces the mean squared error under dual frame designs.

In the EWCTS 2021, the possibility of combining RDD and face-to-face samples was unfeasible due to the COVID-19 restrictions that were in place in the vast majority of Europe. There was also the possibility of combining RDD and landline samples, but due to the increasingly low coverage of landlines, that would have not resulted in a noticeable increase in coverage.

In '3MC' survey design, using the same procedures across countries is not necessary for optimizing comparability (AAPOR 2021). On the contrary, in a multinational survey, according to Kish (1994), sample collecting process and design may adapt to each national resources and its potential to account for increasing probabilities of gathering all population elements. Therefore, the flexibility showed by Eurofound to adapt sampling designs (such as the case of Sweden) is a demonstration of flexibility required for this kind of surveys and is also being done in many others (ESS, American National Election Studies, World Value Survey, European Value Studies).

2.3.4 Sampling plan and sample representativity

Table 17. Indicators in sampling plan and sample representativity

Number	Criteria	Indicator	Target	Assessment
16	Accuracy	Percentage of countries where net sample size >= 1000	100%	Target met
20	Punctuality	Sampling preparation timetable adhered to	Yes	Target met
22	Punctuality	Gross sample available to national agencies in sufficient time to start fieldwork	Yes	Target met
28	Accuracy	Percentage of countries where the net sample size >= planned sample size	Yes	Target met
23	Accuracy	Percentage of countries where the distributions across agreed reference statistics categories of the net sample closely approximates the distributions of the universe (deviations in the proportional size of each of the strata between the two should not exceed 20% or 1 percentage point - whichever is the larger number) - including gender, age category, self-employed, working part time, education level 'low', sector (top-level NACE post-coding), occupation (top-level ISCO post-coding), region, urbanity	100%	Target not met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

Three of the indicators proposed refer to the Accuracy level while the others two refer to Punctuality.

Most of the indicators were fully met. The case of last Accuracy indicator must be noted, given that no country met the target on gender, age and education simultaneously (samples of 34/35 countries met the target on gender; 12/35 met the target on age and 3/35 met the target on education). This indicator shows a poor behaviour that compromises the reliability of the sample. In some countries, women are overrepresented by 10%, while in other countries people between 50 and 74 years of age are underrepresented by 20%. The difference between reference statistics and sample statistics is even greater, given that the gap is above 40% in some countries. This difference may be a consequence of the RDD selection method, which makes it difficult to obtain a sample for which the distribution of sociodemographic variables resembles that of the reference population, given that some population groups (such as wealthy, better educated, younger or middle-aged people) are more prone to have mobile phone access than others (see Pasadas-del-Amo, 2018 for a review on the differences between mobile phone users and non-users and its consequences on the estimation of population parameters). Therefore, it became fundamental to reweight the sample according to the sociodemographic and working characteristics, although this may not completely remove the non-response bias given that it may be driven by other variables where similar gaps were found (such as self-employment or subemployment measures).

The report does not detail the procedure for determining the sample size in each country, but one of the actors of the sampling design procedure confirmed that the sampling followed a compromise allocation, where each country was aimed to have a minimum of 1000 individuals in the sample, and some extra individuals based on the available budget for each country (all of them were able to have larger samples that may allow them to do within-country sub-analyses) and the size of each country.

Determinations of sample size and necessary precision are key issues relating to sample design in '3MC' surveys (AAPOR, 2021), and therefore some details from the procedure to allocate sample size should be included in the sampling report.

2.4 Weighting

As the distribution of groups of observations in this survey dataset differ from the distribution in the target population due to the sampling frame, the sample design, and patterns of unit nonresponse, weighting is one of the best ways to obtain more reliable estimates by reflecting the effect of different selection probabilities on them. The objective of weighting is to reduce the negative effects of nonresponse and out-of-scope problems. Verma (2014) defined the weight adjustment process in five steps: calculate design weights, adjust these weights to compensate for nonresponse, calibrate the weights to known totals obtained from the external data sources, trimming and scaling of the weights. The procedure followed by the contractor in charge of the weighting procedure follows this scheme.

2.4.1 Analysis of quality indicators on weighting

Twenty indicators are included in the QAP which are related to the weighting process. The Accuracy indicators for weighting are related to reference statistics, the basic design weights and post-stratification and trimming weights. Indicators on data Accessibility of several weights are also included, a very important aspect for carrying out subsequent estimates.

Table 18. Indicators on weighting procedure related to accuracy

Number	Criteria	Indicator	Target	Assessment
38	Accuracy	Percentage of countries where the weighting strategy integrates all available information on those elements that are foreseen to be included in the weighting procedure, given the sampling plan	100%	Target met
10	Accuracy	Percentage of the population covered by the reference statistics	100%	Target mostly met
41	Punctuality	Percentage of countries for which the reference statistics by post-stratification variables have been collected.	100%	Target met

Number	Criteria	Indicator	Target	Assessment
44	**Accuracy**	Percentage of countries where the design weight is specified in accordance with the sampling design	**100%**	**Target met**
48	**Accuracy**	Percentage of countries where a common set of variables with common categories are used for weighting	**100%**	**Target mostly met**
51	**Accuracy**	Percentage of countries where the weights are based on up-to-date official population statistics collected within two years preceding fieldwork	**100%**	**Target met**
54	**Accuracy**	Weight trimming follows the weighting strategy and is fully documented and replicable	**Y**	**Not applicable**

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

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. The calibration procedure appears to be comprehensive of all available information for all countries, thus meeting the target. The second has been evaluated as Target mostly only because this indicator was met in all countries except Malta.

One of the main indicators, which is fundamental in '3MC' surveys, is that a common set of variables with common categories were used for weighting in all countries. In this sense, the target percentage of countries using the same weighting variables was not met. The reason was that, in some countries, the calibration variables were slightly different, especially because of the merging of classes: there were countries where, for some variables, (most of the times, those related to occupation) some levels had to be merged because of not having enough sample size or a very small population size for some of them. For example, in some countries the number of farmers was too low, so that sector had to be combined with another one, or in some Balkan countries the number of males over 65 years old was too low, so the stratum was combined with females over 65 years old.

Differences among countries in the type and quality of external data available for post-stratification adjustments are not uncommon and take place in other '3MC' surveys (Eurofound, 2016). For the EU Member States, all the statistics used for weighting were obtained from Eurostat. For non-EU countries some variables could not be coded exactly as in EU countries. The problem with the UK must be noted: the UK LFS variables are no longer harmonised with Eurostat in important variables such as age, economic sector groups and occupation groups. This is an issue that will likely keep happening in future editions of the survey, among other comparative official statistics.

Regarding the weight trimming, it must be noted that the weights have finally not been trimmed because the calibration method was bounded linear calibration, which works with boundaries in the g-weights so there is no necessity to trim the weights after their calculation. Therefore, the indicator is not applicable in this case.

Table 19. Indicators on weighting procedure related to accessibility

Number	Criteria	Indicator	Target	Assessment
39	Accessibility	Percentage of countries for which the weighting strategy and procedure are made completely transparent in the weighting report	100%	Target met
45	Accessibility	Design weight included in the dataset	Y	Target met
46	Accessibility	Procedure for constructing design weights outlined in sampling report	Y	Target met
50	Accessibility	Procedure for constructing post-stratification weights outlined in weighting report	Y	Target met
52	Accessibility	Supra-national weights included in dataset	Y	Target met
53	Accessibility	Procedure for constructing of and sources used for supra-national weights described in weighting report	Y	Target met
55	Accessibility	Trimmed and untrimmed weights are included in the dataset	Y	Not applicable
56	Accessibility	Trimming cut-off points and number of trimmed cases for each country are included in the weighting report	Y	Not applicable

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

The reports include detailed information about the weighting procedure and the steps followed to do the adjustments. The weight trimming indicators are not applicable for this survey, given that the weights have finally not been trimmed because the calibration method was bounded linear calibration, which works with boundaries in the g-weights so there is no necessity to trim the weights after their calculation.

2.4.2 Selection of variables for reweighting

The variables used for calibration remained the same as in previous editions of the EWCS. They are basic sociodemographic variables: age, sex, region, occupation, and economic sector. As explained above, they do not keep equal categories across all countries as some of them had to be combined for sample or population size reasons, which makes the process more difficult and could add some variability in the calibration procedure. It could also affect the cross-comparability if the nonresponse patterns are different, although we do not expect this effect to be large.

On the other hand, these variables could not be sufficient to explain and reduce coverage and nonresponse biases. The choice of calibration variables is always limited by the amount of information available, especially in a working conditions survey such as this one, where the target population is constantly changing, and it is not usually considered as a sociodemographic group in censuses or other population statistics. However, the response could be driven by other variables such as the type of employment; the sampling report shows some gaps in job-related variables which could not be shortened by weighting on the sociodemographic variables. Including these variables in the reweighting procedure could contribute to reducing nonresponse bias, but on the other hand this could also increase the risk of introducing comparability errors.

Selecting the calibration variables is a complex task. Research on variable selection for predicting the response probability (propensity score methods) shows that bias reduction adjustments can improve if prognostically important variables are used, this is, variables that are related both to the selection mechanism and to the variable of interest (Hirano and Imbens, 2001; Brookhart et al., 2006; Austin, 2008). This may also be the case for calibration adjustments, but the information about the predictive power of the variables may be limited in most of its applications. For this reason, some statistical procedures have been developed to select variables according to the available data. Silva and Skinner (1997) and Clark and Chambers (2008) developed methods based on minimising the mean squared error of the prediction using stepwise variable selection. More recent approaches consider the least absolute shrinkage and selection operator (LASSO) regression to select variables for calibration (Chen, 2016; Tsung et al., 2018; Chen et al., 2019). The implementation of these methods also allows calculations of the variance of the estimators with analytical expressions.

2.4.3 Checking the reweighting method.

The original design weights were calculated for all countries as the inverse inclusion probability of a simple random sampling without replacement (SRSWOR) scheme: number of employed people divided by size of the sample. The sample design, using RDD in all the countries except for Sweden, should constitute a SRSWOR scheme in practice, and therefore this approach is valid. However, the sampling report specifies that a stratified sampling design (by age and sex) was used for Sweden. If the sample allocation was completely proportional to the size of each stratum, the estimators are the same as in a SRSWOR scheme, and therefore the procedure for calculating design weights should be valid as well, but small deviations from the theoretical proportional allocation (which could happen as a result of rounding or fieldwork issues) could result in incorrect design weights. The sampling report mentions that the deviations from the theoretical proportional allocation should be small, meaning that this should not be a very important issue for the results and the quality is therefore ensured.

The second adjustment is multiplying the design weights by an adjustment factor that is equal to the inverse number of mobile phones owned by each respondent. This is a fair adjustment, as people with more mobile phones may have a greater probability of being selected in a sample collected using RDD. This adjustment is not applied to the Sweden sample as it is selected directly from registers. However, the adjustment factor is capped at 2 mobile phones: if someone owns 3, 4, 5 or more mobile phones, their adjustment factor is the same as if they owned 2 mobile phones. This means that people with 1 mobile phone have their design weight multiplied by 1, while people with 2 or more mobile phones have their design weight multiplied by 0.5. This cannot be assessed from the sampling report, which reads in the main text that the adjustment factor was capped at 4 phones, while the supporting formula indicates that the adjustment factor was capped at 3 phones. We could confirm that the cap was on two mobile phones after a brief exchange with one of the actors involved in the weighting procedure.

Both the sampling report and the interviewees involved in the sampling and weighting procedure cited variance reduction reasons: introducing an adjustment factor may reduce bias but also increase the variance, which could be counterproductive. One of the interviewed argued, however, that the capping of the number of phones should have been assigned on a case-by-case basis, studying the distribution of the number of phones in each country and capping at a very high percentile according to that distribution. On the other hand, an independent expert in survey sampling and weighting expressed some concerns in an interview regarding this capping procedure, considering that the primary objective of weighting is to reduce total error.

In the last step prior to calibration, the adjusted weights are rescaled to make their sum equal to the population size: they are multiplied by N (the number of employed people) and then divided by the sum of the adjusted weights. It is easy to prove that this scaling makes the weights strictly dependent on the number of phone numbers of a given person:

$$adjustment\ factor\ (scaling)\ =\ w_{gross} * \frac{N}{\sum_{s} w_{gross}}$$

$$= \frac{N}{n} * adj(phones) * \frac{N}{\sum_{s} \frac{N}{n} * adj(phones)} = N * \frac{adj(phones)}{\sum_{s} adj(phones)},$$

Where

$$adj(phones) = \frac{1}{(min\ (number\ of\ mobile\ phones\ ,2))}$$

Given that the only possibilities for $adj(phones)$ is to be 1 or 0.5, the final formula can be expressed as follows:

$$\sum_{s} adj(phones) = n * \alpha + \frac{n}{2} * (1 - \alpha) = n * \left(\alpha + \frac{1}{2} - \frac{\alpha}{2}\right) = \frac{n}{2}(1 + \alpha),$$

where α represents the proportion of people in the sampling frame who only owns one mobile phone. It can be proven that the variance of these weights reaches its maximum when α = 0.333, while decreasing steadily towards zero as the proportion α grows. Given that, in 2021, the mobile cellular subscriptions in the European Union were 123 per 100 people (World Bank, 2021), we consider that is highly unlikely to have such a low value for α and more feasible for it to be above 70%, given that a number below 70% would result in a number of subscriptions per capita larger than 1.23 (even considering that the rest of the population only has two phones). In such a situation, leaving the number of phones uncapped would affect only a very limited part of the sums and variances above, meaning that it should not cause a very noticeable effect in the variance.

On the other hand, if we consider the number of phones, the actual inclusion probabilities of everyone in the population should be:

$$\pi_i = \frac{n * number\ of\ phones_i}{N * mean\ number\ of\ phones}$$

If we consider the weights as the inverse of the inclusion probability,

$$w_i = \frac{N * mean\ number\ of\ phones}{n * number\ of\ phones_i}$$

However, after the previous adjustments, the (uncapped) weights are

$$adj(phones)_i = \frac{N/number\ of\ phones_i}{\frac{n}{2}(1+\alpha)} = \frac{N}{n * number\ of\ phones_i} * \frac{2}{1+\alpha}$$

Even if we consider a population where people only have 1 or 2 mobile phones, the mean number of phones would be $\alpha + (1-\alpha) * 2 \; = \; 2-\alpha$, which results in a difference between weights of the following order:

$$\begin{aligned} adj(phones)_i - w_i &= \frac{N}{n * number\ of\ phones_i}\left(\frac{2}{1+\alpha} - 2 + \alpha\right) \\ &= \frac{N}{n * number\ of\ phones_i} * \frac{\alpha^2 - \alpha}{\alpha + 1} \end{aligned}$$

The difference between the estimated and the actual weights reaches a maximum when $\alpha = \sqrt{2} - 1 \approx 0.41421$, although still noticeable differences can be found with values of α around 70-80%. If we consider a population with people with more than 2 mobile phones, the differences are expected to be larger. This induces a bias in the weights that might be larger as the sample size decreases.

2.4.4 Nonresponse adjustments

The non-response adjustment is an important step in a survey where the non-response rate is as high as in the EWCTS 2021. A reweighting strategy was adopted, using the design weights calculated in the previous stage (described in the previous section), which is an adequate strategy as the high non-response rate would make it very difficult to adopt other strategies such as imputation or substitution. All the reweighting procedure was done using the version 4.0 of software Calif, developed by the Statistical Office of the Slovak Republic. This is a software that offers a user-friendly interface for doing calibration while providing a set of useful tools and visualizations, such as histograms and statistical summaries of weights. We consider this software to be more than acceptable for the task conducted in the reweighting procedure.

For the weighting, Eurofound chose the linear bounded method and the calib solver. The linear method uses what is known as the quadratic distance function, proposed by Deville and Särndal (1992). This is a standard method for calibration that has been widely used but has several drawbacks such as the possibility of providing negative weights or that the method assumes a linear relationship between the variable under study and the variables used in the calibration. This assumption is difficult to assume when the response variables are categories, as is frequently the case in this survey. Eurofound decided to use the linear bounded method, which imposes the g-weights to be between certain limits. These limits can be used to impose the weights to be above 0 and therefore avoid negative weights, while on the other hand can contribute to reduce the variance as they can be used to avoid too large weights. This procedure is much more recommendable than trimming the weights afterwards, as the weights are optimized to be consistent with the calibration equations when remaining in each interval, while trimming afterwards would require readjusting weights to match the population totals and could result in suboptimal weights. However, the bounds must be chosen carefully to avoid undesirable effects; the independent expert interviewed for this quality report pointed out that bounds, although useful, should not be too stringent to give some flexibility.

The sampling report mentions that the bounds were decided on a country-by-country basis, and the interview with one of the interviewed of the weighting procedure confirmed that this decision was taken by checking the visualizations provided by calif (such as histograms or boxplots) and the Average

Feasibility Difference index, which should not be below 0.7-0.9. They also confirmed that the lower bound was much more important to the weight properties than the upper bound, which is also noticeable when looking at the histograms of the weights provided in the sampling report: in many countries, the maximum weights are nowhere near the upper bound. In addition, the distribution of weights does not seem to be centred around zero in many countries; it seems to adjust better to a log-normal distribution, meaning that it is important to set the lower bound, but it should not be completely relevant regarding the final estimates.

The reweighting procedure was done in three steps: first, age and sex were included as calibration variables. Then, the region variable was added, and finally occupation and sector were added in the third step. The multi-step calibration is a widely studied topic in literature (Kott and Chang, 2010, Kott, 2016, Singh and Sedory, 2016), but it requires a specific framework for estimation which is not possible to know whether it was applied in this case or not. More information about this three-step procedure should be given, along with the reasons to use this approach; we assume that it was done to avoid too many calibration totals at a time (which could be troublesome regarding calibration convergence), but it should be clarified in the report.

The treatment of small or empty cells was done by merging adjacent cells in the population in those variables where needed. This was indeed true for the age variable, where small cells were merged with adjacent age groups, or unified within all genders (for example, instead of having one cell for "female over 60" and another one for "male over 60", they would have a single cell for "female and male over 60"). This was also true for the region, where the merging procedures only concerned adjacent regions, and some other regions were split. Regarding occupation, the two sections that were merged (where needed) were skilled agricultural workers and artisans; the similarities between the people in the two sections is not completely clear but it is assumed that there was no better choice, given the variable characteristics and the disparities between artisans and the rest of occupations. Finally, the sectors that were usually merged (where needed) were 'R' (arts entertainment and recreation) with the combination of 'S' and 'U' (other service activities and activities of extraterritorial organisations and bodies, respectively), which could be adequate assuming that the latter sectors would already have a considerable level of heterogeneity. Other merging procedures included sectors A, B, D and E, related to agriculture forestry, mining, electricity and water supply and waste management. Most merging procedures took place in countries with small sample sizes, such as Balkan countries or Switzerland where the sample sizes were around 1000-1200. As mentioned earlier, these procedures cause the calibration variables to be different across countries, which could induce comparability problems in the final estimates.

2.4.5 Reweighting for the variables included in the randomization modules

An important detail is that some questions were not asked to all the individuals, but to different groups allocated randomly: these questions were asked only to two thirds of the sample, meaning that a third part was lost for them. Given that the allocation is random, the sampling reports that this procedure should not have any consequences on sampling bias and would only result in larger sampling errors due to reduced sample sizes. However, it is not clear if facing different sets of questions could influence respondents to break-off before ending the interview, meaning that non-response mechanisms could be different across groups.

The weighting and sampling report mentions that it is unclear which weights were used for certain observations, potentially causing imprecise estimates. It suggests that the same weights were likely used for all respondents, leading to less accuracy. The proposed solution involves a two-phase sampling approach to calibrate the weights, but the report lacks detailed information on this process.

Another issue related to modularisation is that the real sample size is not clear for the variables that have been assigned to these modules. According to the report, "Additional sample was released in several countries to achieve additional completes due to an issue in the modularisation design and allocations" (Eurofound, 2021d), but the issues are not detailed nor the sample size that was missing in those modules. The issue was related to quality issues during the implementation in which the automated allocation mechanism of the planned missingness design failed as previously discussed in the fieldwork section. However, it is assessed that the number of valid responses to the modularised questions (job quality questions) corresponds to approximately two thirds of the sample size of each country, using the raw data.

2.4.6 Weight analysis

The sampling report includes a very detailed study about the final weights, including the extent of the adjustment that had to be done in the calibration step (measured as the difference between population totals and estimated totals from the survey prior to calibration using design weights), a comparison between unweighted and weighted estimates from the survey and estimates from the LFS of each country in a number of monitoring variables (allowing for a bias analysis), and an analysis of the design effect in each country. This study is largely insightful, and some important conclusions arise from it:

- If using only design weights, the sample tends, overall, to underrepresent elderly people (specially in Balkan countries) and overrepresent younger people, except for some Nordic countries where the opposite applies. It also overrepresents higher skilled occupations and underrepresents "blue collar" ones (agriculture, mining, manufacturing...). Finally, it also tends to overrepresent the population from urban areas, especially those from the capital cities of each country. These phenomena are largely common in telephone surveys (Pasadas-del-Amo, 2018).
- The sample also provides biased estimates for some of the monitoring variables, which weighting adjustments can remove only to a limited extent, because the biases associated to some of those variables are unrelated to the calibration variables. In the latter cases, the results presented in the report shows that calibration does not produce very important changes in the estimates. This is a desirable property, given that, if the weights are ineffective for removing bias, they should at least do not contribute to increase it or the variance. However, it also shows that the calibration variables used could be not enough for some of the potential variables of interest.
- The design effect is very noticeable overall, especially in Balkan countries, where the effective sample sizes are almost half of the actual sample size. As cited in the report, this is a common phenomenon in this kind of surveys, and in this edition the design effects are even lower for some countries than in 5th edition.

To check some of the conclusions above and other issues of interest, we have conducted a short analysis of the final weights of the sample. In a first step, we have summarised the distribution of weights for each country; the results can be found in Table 30 and Figure 3.

The disparities across countries are noticeable: individuals from smaller countries, where even a modest sample size can be considered relatively large when compared to the country size, have smaller weights than individuals from larger countries, where everyone of the sample is ought to represent a larger group. The median Maltese respondent represents approximately 182 individuals with their characteristics, while the median German respondent represents approximately 10,046 individuals. Such disparities can increase the variance of the final estimates for Europe as a whole, in comparison to a situation where the sample sizes are much more balanced according to the population size of each country. On the other hand, optimizing the allocation of the samples according to the proportional criteria, as long as the sample size for the whole European Union is not increased, would lead to other kinds of issues such as having small sizes in smaller countries, which would have a dramatic impact on the estimation of their main variables, while the positive impact in larger countries would not be as larger as the negative impact in smaller ones.

In a second step, we have checked whether the sum of weights equals the target population or not. We have done this for each country, for both the whole target population (number of employed people) and age and sex strata combined. The differences are calculated in absolute and relative terms, with the following formulas:

$$Absolute\ difference\ = \sum_{\square}^{\square} \square\, calibration\ weights - N\ (number\ of\ employed\ people)$$

$$Relative\ difference\ = \left(\frac{\sum_{\square}^{\square} \square\, calibration\ weights}{N\ (number\ of\ employed\ people)} - 1\right) * 100$$

A large absolute difference may not be the cause of a deficient calibration procedure but a matter of scale, while a large relative difference may be more informative of possible discrepancies in the calibration procedure. On the other hand, the absolute numbers give an idea of how the difference may affect the final estimates for the whole of Europe.

The number of employed people has been obtained from Eurostat figures from the year 2021 (when the survey took place). To be consistent with the sampling report, we have used the indicator [lfsa_egaps]: Employment by sex, age and professional status. This is the same indicator that was used for obtaining sample profile sort. However, we did not consider the professional status given that the definitions of professional status in the survey are not fully consistent across countries. The operation therefore consists in obtaining the sum of weights for individuals in each combination of age and sex, and comparing this sum to the number of individuals in the population of employed people that belong in the combination (of age and sex) according to the LFS of each country. However, there are some limitations for the data source; more precisely:

- Data from the United Kingdom was discontinued in 2019. The sum of weights of the respondents from United Kingdom has been compared to the figures from 2019.
- 2021 data from North Macedonia was not available; a report from North Macedonia State Statistical Office could be retrieved (MAKSTAT, 2023), but the figures of number of employed

people in 2021 were largely different to those reported by Eurostat from 2020 (693,464 employed people in 2021 vs 794,400 employed people in 2020). For this reason, and as a matter of consistency, 2020 data from Eurostat was used instead.

- 2021 data from Montenegro was also not available in Eurostat but could be retrieved from the report on the LFS done by the Montenegro Statistical Office (Monstat, 2022), whose figures were compatible to those provided by Eurostat for previous years. We consider this figure to be consistent, and therefore it was used for the calculations.
- Data from Albania, Bosnia & Herzegovina and Kosovo could not be retrieved.

The results for the whole target population of each country can be observed in Figures 4 and 5.

It can be observed that the sum of weights is consistent with the number of employed population in each country, with a few exceptions (notably Germany) whose figures are slightly off, but these differences can be considered negligible. In the case of Germany, it must be noted that this analysis has been carried out after some revisions in the source data from Eurostat, meaning that the difference of 100,000 workers could be an effect of further corrections and therefore could not be avoided at the time the calibration was done.

The absolute and relative differences for age and gender strata can be observed in Figures 5 and 6. The age was cut into three intervals: 15-24 years old, 25-49 years old and 50-74 years old. The individuals from the sample that belong to each interval were obtained manually by discretising the variable [age] of the dataset. This was done because these were the only age intervals provided by the [lfsa_egaps] indicator from Eurostat. Some individuals older than 74 years old or with no age data were discarded for the analysis. On the other hand, people who did not identify with male or female options in the [gender] variable were statistically treated as male, following the same principles that were used in the sampling and weighting procedure according to the sample report (See Figure 6). It can be noticed (See Figure 7) in these figures that the largest differences between the expected and the actual sum of weights take place in younger strata, while the weights for people between 25 and 49 years old fit very well with the population totals. The differences for people between 50 and 74 years old are somehow larger but not as large as those for younger people. It does not seem that the gender plays a role, as the differences remain very similar across genders. The absolute differences show that the large differences observed in younger people may be an effect of smaller denominators, as the gaps are rarely above 20,000 units. However, these differences may not be negligible, especially when doing estimations for younger branches. It must be noted that the EWCTS 2021 do not have respondents aged 15, meaning that the comparison in the younger age group could be missing a certain portion of workers, which could explain part of the gaps observed for these groups, and therefore the gaps would not be a weighting issue but a limitation of this analysis.

In a third and final step, a regression analysis was performed to assess the reasons behind the differences in design effects across countries. For the matter we obtained, for the respondents from each country, the median age, the proportion of women, the proportion of higher skilled workers (those with occupations whose ISCO-08 code was below or equal to 5) and the proportion of “white-collar” workers (those with occupations whose NACE Rev. 1 label was J, K, L, M or N). In an initial step, the model also considered the sample size of each country as an explanatory variable, but it turned out to have no predictive power and provided worse values for goodness-of-fit indicators (AIC and

Adjusted R2 coefficients). These were the variables involved in the calibration procedure (apart from region, whose comparability would be troublesome) and therefore expected to be tied to the design effect. The regression model proposed to explain the design effect of the i-th country was the following one:

$$deff_i = \beta_0 + \beta_1 * Median\ age_i + \beta_2 * Proportion\ of\ female_i + \\ \beta_3 * Proportion\ of\ higher\ skilled\ workers_i + \beta_4 * Proportion\ of\ "white - collar"\ workers_i$$

We also considered a model with interactions that provided a larger Adjusted R2 but discarded it because of its low interpretability. The results of the model above can be consulted in Table 31.

It can be observed that the model has a poor predictive value and that the only significant variable is the median age. If we consider the bivariate relationship between design effects and median age, the Pearson correlation coefficient is -0.5045 (R2 of 0.2545 for a simple regression model with only the median age as independent variable). This is the largest correlation between the design effects and any of the four variables included in the model above, which shows that the design effects were mainly driven by the age. In fact, the correlation shows that the younger the median age, the larger the design effects were. A possible explanation could be that underrepresenting the elderly and overrepresenting the youth results in a lower median age, but also results in more stringent adjustments to correct the age bias which increase the design effects.

2.4.7 External sources used in reweighting.

The calibration for some variables is not done on known population totals but on the LFS estimates for each country. These estimates are based on large-scale samples with a curated sampling design, and therefore the figures should be accurate, but in the end, they are estimates and may have sampling errors. When the population controls do not exist or cannot be found, many researchers use survey-estimated control totals, and apply traditional variance formulae as if the controls were known without error. The sampling report (Eurofound, 2021d) accounts for this drawback, and a simulation study can be found on the influence and magnitude of LFS errors in a couple of countries. However, the authors assume SRSWOR for the LFS, but the actual sampling design is stratified. This kind of design can increase the variability of estimates if the strata are not very different between them, and therefore the range of estimates could be even larger.

An independent expert, consulted for this report, mentioned in a personal interview that the impact of using estimates instead of actual population totals depends on the situation. It is a procedure that could induce bias if the estimates are biased, but that should not be the case of LFS. Regarding the variance, the key aspect is the relative sample size: if the LFS sample size is much larger than the EWCTS 2021 sample size, the increase in variance will be negligible. If both sample sizes are closer (or even equal), the increase in variance will be much more noticeable. In this case, the former situation applies, but it could be assessed via bootstrap variance estimators.

Opsomer and Erciulescu (2021) discuss this problem and propose several methods to apply calibration weighting adjustment to full-sample weights and to each column of replicate weights. These statistical methods differ in the way they generate different control totals for each column of replicate weights and in the type of data they require the analyst to use. The method of Fuller (1998) requires the analyst to have a variance-covariance matrix for the estimated control totals, while the method of Opsomer and Erciulescu (2021) requires the analyst to use the full dataset for the control survey along with associated replicate weights. Both methods are implemented in the 'svrep' package (Schneider, 2023). This issue should be mentioned in the estimates and the use of such techniques to variance estimation should be considered or suggested in the sampling report.

Overall, our quality assessment of the weighting concludes that the EWCTS 2021 has followed sound principles for its sampling design and weighting procedure. The sample has been carefully designed to be comparable across all the participant countries, the sampling frames are relatively wide regarding coverage, the sample sizes are large enough to produce reliable national estimates and the fieldwork has taken place without major issues.

The weighting system has been implemented following regular standards used in calibration, with a proactive construction of design weights taking over coverage into account, a calibration procedure in various steps to avoid further problems, using auxiliary variables that may have correlations with potential variables of interest, and using linear bounded distances which avoids further weight trimming. In addition, the analysis of the weighting procedure has been well undertaken in the report.

3. Survey Outputs Quality Assessment

The framework adopted to assess the quality of the outputs is the Total Survey Error (TSE) (Groves, et al. 2009) a paradigm previously detailed, which helps organise and identify error sources and estimate their relative magnitude, which can assist those planning '3MC' surveys (Johnson, et al., 2018) and to evaluate design and implementation trade-offs. In this case, the assessment concerns the quality of the microdata and paradata obtained in the survey considering some important aspects related to internal and external validity.

The results are first analysed and discussed in relation to their compliance with EWCTS 2021 Quality Assurance and Control Report (Ipsos, 2021a), making a set of suggestions relevant for upcoming EWCS surveys. In addition to the evaluation report, the evidence provided in Technical Report (Ipsos, 2021c), Data validation and editing report (Eurofound, 2021c) for the EWCTS 2021 and all the documentation before and during the fieldwork provided by Eurofound, is reviewed. Technical reports from previous waves have been used also consider.

The QAP includes 8 indicators related to microdata (146-153) and another 8 related to paradata (154-161). In 6 of the indicators related to microdata, the quality criterion is related to Accuracy, 1 is related to Accessibility and 1 with Punctuality. Regarding the indicators related to the paradata, 7 of them are related to the Accuracy quality criteria and 1 with Punctuality. Although the quality indicators have five dimensions, our assessment focuses primarily on the indicators related to the dimension of Accuracy, which is considered fundamental to product quality (Biemer et al., 2014). Biemer and Lyberg (2003) viewed accuracy as the dimension to be optimised in a survey and they argued that sufficient accuracy is essential for the other quality dimensions to be relevant.

3.1. Microdata and paradata

3.1.1 Analysis of quality indicators on paradata

Producers of '3MC' survey data should facilitate available paradata (AAPOR, 2021). The data set provided by Ipsos at the end of the mainstage fieldwork was a combined data set of paradata, survey data and quality control data.

In this case, a file is observed that includes detailed paradata information related to the country, sample frame, number of contacts, a summary of weekend/weekday call attempts, fieldwork period, last outcome status, interviewer ID, interviewer gender, interviewer age, interviewer education, interviewer language, outcome code, call time and contact status of each of the 50 call attempts.

Table 20. External evaluation paradata indicators

Number	Criteria	Indicator	Target	Assessment
154	**Accuracy**	Percentage of variables that are named in accordance with agreed template	**100%**	**Target met**

Number	Criteria	Indicator	Target	Assessment
155	Accuracy	Percentage of variables that are labelled in accordance with agreed template	100%	Target mostly met
156	Accuracy	Percentage of variables for which the missing values are properly defined	100%	Target mostly met
157	Accuracy	Percentage of variables for which the level of measurement is properly defined	100%	Target mostly met
158	Accuracy	Percentage of para data variables included in the dataset, as agreed with Eurofound	100%	Target met
159	Accuracy	Percentage of stratification variables included in the dataset	Y	Target met
160	Punctuality	Paradata delivered at agreed date	Y	Target met
161	Accuracy	Dataset delivered in specified format	Y	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

These indicators are mostly adequate to give an idea of the quality of the data provided. Some of the indicators have been qualified as mostly met since some minor issues have been detected and detailed in Annex 5.

3.1.2 Analysis of quality indicators on microdata

Table 21. External evaluation microdata indicators

Number	Criteria	Indicator	Target	Assessment
146	Accuracy	Percentage of variables that are named in accordance with agreed template	100%	Target met
147	Accuracy	Percentage of variables that are labelled in accordance with agreed template	100%	Target mostly met
148	Accuracy	Percentage of variables for which the missing values are properly defined	100%	Target mostly met
149	Accuracy	Percentage of variables for which the level of measurement is properly defined	100%	Target mostly met
150	Accuracy	Percentage of substantive variables included in the dataset	100%	Target met
151	Accessibility	Datasets delivered in specified format	Y	Target met
152	Punctuality	Substantive datasets delivered at agreed date	Y	Target met
153	Accuracy	Syntax to create the trend data file checked by Eurofound	Y	Not applicable

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

A draft template for the dataset was prepared and agreed and revised accordingly prior to the fieldwork in 2021. This ensured that the data conversion could be prepared to match the final delivery requirements. Thus, most of the indicators were fully achieved.

The last Accuracy Indicator was not met, but according to the contractor it was agreed that it was not necessary to provide this since the EWCTS 2021 data was not destined to be included in the EWCS (face-to-face) trend data file.

Indicators 147, 148 and 149 were almost fully met, but we have qualified them as mostly met because some minor issues and points were detected that have been commented on Annex 5.

3.1.3 Assessment of the quality of the data

In relation to the microdata, it must be considered that although many of the survey questions are also contained in previous versions of the EWCS, the EWCTS 2021 has several differences in its methodology. All interviews were conducted over the phone instead of face-to-face, so the questionnaire had to be adapted for phone interviews.

Some of the results are hence likely to be affected by changes in methodology, as well as changes in working conditions. For this reason, direct comparisons with previous editions of the EWCS may not be possible for some variables (variables that refer to labour aspects that are related to the pandemic). If comparisons are made, the different methodology must be considered in the analysis.

On the other hand, the use of paradata to investigate and reduce survey error provides a wide range of information about the survey data collection process. Survey errors are especially important in ‘3MC’ surveys due to their comparative nature. Paradata provides an additional tool to evaluate and reduce survey error sources across participating countries (Kreuter, 2013; Kreuter and Olson, 2013). The use of a standardised CATI instrument in the EWCTS 2021 facilitated a standardised collection of paradata. The dataset, containing the paradata, was prepared and agreed upon prior to the main fieldwork in 2021, and includes enough information to find out if the fieldwork rules were followed, if they were followed well, and identify issues, like issues in the interviews.

3.1.3.1. Internal validity

Internal validity refers to the extent to which the concepts are measured accurately and precisely. Assessing measurement error is challenging because there is no ‘gold standard’ against which the response can be compared. Some important aspects that affect this precision are considered.

Sample composition statistics

A descriptive analysis was performed, including the main statistical measures of central tendency and shape, as well as histograms, boxplots, and detection of outliers of all variables both for sociodemographic variables and relevant variables (Annex 1 for qualitative and Annex 2 for quantitative variables).

Coefficient of variation and design effects

The precision of a measurement is easier to assess as it relates to the sampling design. Some indicators of the precision of a survey are the standard error and the design effect, which is reported in the sampling and weighting report.

Standard errors depend on the sample size and sampling design as well as the specific variable that is being considered. We calculate the standard errors, the coefficient of variation of some relevant variables in general and in each country, considering the weights and the design of the survey.

Table 37 shows some of the variables of interest together with their respective errors. To see the values of the variables by country see Annex 3. In general, the variables studied have a small dispersion~~.~~

The design effects provide an indication of the additional variance introduced by the weighting procedure. Design effects for each country are provided in Table 38. There are large differences between countries. In some countries it is around 2, dividing the power of their samples in half. The values of the design effect are in a similar range to those of the 2015 survey~~.~~

To analyse the differences between the effects of the design between countries, a regression analysis was carried out, which is shown in the section 2.8.5 Weight analysis. This analysis shows a low predictive capacity of the model, and low correlation between the variables included in the model.

Nonresponse errors

Unit nonresponse

An important source of bias in survey estimates is non-response (Groves et al, 2001). Adjustment for the effects of unit non-response bias is typically made by weighting based on population data. Such approaches can only ever correct for that proportion of non-response bias that is explained by the weighting classes. They therefore rely upon an assumption of strong correlation between the classes and the survey measures, as well as requiring correlation between the classes and response propensity (Lynn, 2002). In the EWCTS 2021 the response rate was 5%, thus the potential size of non-response bias is quite high. The problems of non-response that the RDD method used entails are clearly manifested.

To better analyse non-response, it is useful to calculate the different response rates, according to Aapor's coding (AAPOR, undated). Ipsos and Eurofound agreed on general grouping principles for the outcome codes. Some measures that can give us an idea of the nonresponse behaviour in this survey have been calculated. Figure 8 shows the yield rates (final number of interviews achieved after all quality checks/actual gross sample) for the different countries. A great variability is observed: the minimum value is 0.095%, which corresponds to Germany and the maximum 11.616% which corresponds to Bulgaria. The median is at the point 5.250% and the first and third quartiles at 2.976%, 6.771%, thus having an interquartile range of 0.038.

Based on the data of the people who answered, and on the reference statistics obtained from Eurostat (Ipsos, 2021a), an important variability can be observed by countries in the representation by sex: in some countries the overrepresentation of women exceeds 10% (Estonia) while in others there is an underrepresentation of women of -7% (Albania). There are also distortions in the representation of age groups (over-representation of people younger than 25 was reported in Kosovo +11.08%) and under-representation of young people was reported in Albania (-8.92%), the older population (50-74 years old) is most represented is in Sweden (+14.69%) and the lowest representation for the older population is observed in Kosovo (-20.88%). Issues are also observed in the representativeness of the sample by education level: respondents with a lower education were generally more difficult to reach or convince to participate.

It is clear, then, that the behaviour of non-response by country is very diverse and cannot be explained solely by these three characteristics. As we do not have data on people who did not respond, we cannot model the probability of responding with techniques such as propensity score weighting or quasi-random response models (Kott, 2005, Haziza and Lesage, 2016, Lee 2006, Lee and Valliant 2009, Lee et al. 2010, Da Silva and Opsomer, 2009) that could provide more information on this phenomenon in order to correct or ameliorate it.

The adjustment made dealt with non-response bias through calibration by sociodemographic variables (age and sex, region, occupation, and economic sector) based on survey-estimated control totals (see section 2.8.3 for details on the weighting process). It would be convenient to have information on the original weights of the design in the data file so that any researcher can try another weighting with more relevant and extensive information that is available in each country (since it has been verified that the patterns of non-response follow different patterns by country) in order to obtain more precise estimates of the parameters in each country. The use of the same variables for calibration in all countries is not the best way to obtain estimates (Beaumont 2008, Haziza and Beaumont 2017).

Item non-response

A study of partial non-response was conducted, calculating the percentages of missing values in the entire population and by country for some of the relevant variables. Graphs with the non-response separated by sex and other sociodemographic characteristics are presented in Annex 1 and 2 (see Annex 1 for qualitative and Annex 2 for quantitative variables).

To calculate the non-response, it is necessary to consider the modularization and the dependence of some questions on others. In the case of qualitative variables, we found a big number of missing values in some variables (see an example in the Table 31). Other graphs addressing this issue can be found in Annex 1.

In the case of quantitative variables, the percentage of missing values can be observed in Table 32. The rest of the graphs can be found in Annex 2.

Considering the variables studied in previous points, non-response by country is studied (Table 33 and Table 34). The observed percentage of non-response at country level is quite high. According to the information collected in a qualitative interview, “contract_duration_weeks” and “contract_duration_days” variables are recoded by the analyst into “contract_duration” and “contract_duration_new” to unify the time-units. After calculating non-response for them considering the same dependencies as for the rest from “contract_duration_weeks” and “contract_duration_days” non-response rates of around 58% are found instead of the previous 89%. The lowest non-response rate is found in Albania for qualitative variables with a value around 73%.

3.1.3.2. External validity

External validity refers to the extent to which findings from this survey are similar to findings from other surveys or administrative datasets that measure the same concepts.

There is no gold standard to assess the external validity of ‘3MC’ surveys. A possible strategy is to compare key indicators from the EWCS 2021 to the same indicators from the previous waves. However, it is unclear whether these differences are caused by data quality issues, by changes of the methodology or caused by real changes in the work conditions between 2015 and 2021, especially since the COVID epidemic considerably affected working conditions of a significant part of the population.

To check the external validity, the estimates obtained from the sample has been compared with the only external source from which we have similar data, which is Labour Force Survey (LFS). Figure 9 shows a comparison by country of the values obtained in the European LFS and the values estimated in the EWCS2021 for the means of the variables: Average number of usual weekly hours of work and employed persons working at nights.

Looking at the values of the estimates it can be seen that for the “usual_hours_week” variable they are quite similar, and the correlation between the EWCS estimation and the LFS data across countries is high (0,93). However, for the “employed persons working at nights” variable there are differences for some countries and the correlation is low (0,27). In this variable, the LFS analyses “Employed persons working at nights as a percentage of the total employment” and in the EWCTS 2021 the night variable is “How often do you work at night, for at least 2 hours between 10.00 pm and 05.00 am?” so those differences might be due to the approach of the question and any direct comparison would indeed have measurement error.

In conclusion, once the data is collected, the survey data and paradata is stored in a data file available to the public. The whole process of validation overview, implementation, results, and corrective measures is very well documented in the Data validation and editing report. The final product is a file whose data has a reasonable internal validity. The dataset can be used to compile reliable and accurate statistics at the EU-level. More care is warranted when comparing country-level statistics or when using the data to rank countries because the precision of the estimates varies by country.

The use of paradata to investigate and reduce survey error provides a wide range of information about the survey data collection process. Paradata have an important role in the construction of responsive designs to minimize nonresponse bias and cost as proposed by Groves and Heeringa (2006). However, few if any ‘3MC’ surveys are using responsive designs (AAPOR 2021). Paradata is particularly useful to monitor nonresponse bias in ‘3MC’ surveys (AAPOR 2021) and can be used for calculation of response rates and compare field efforts across countries as in the ESS (Stoop et al 2010). The data file in this survey includes a variety of important process data that can be used to optimize calling strategies in subsequent waves. It can be used, for example, to find the best time to call or determining how many call attempts to make. Ipsos report uses various paradata in analyses of nonresponse bias although a more in-depth study could be carried out to characterise the probability of answering based on the characteristics of number of contacts, fieldwork period, call time as well as gender, education, language and age of the interviewer, for example through multilevel models (van de Vijver and Leung, 1997) or machine learning models (Kufner et al 2022). An example of the use of paradata in non-response analysis can be seen in the Labour market change European Company Survey 2019 (Eurofound ,2019).

3.2. Reporting and dissemination

The Reporting and Dissemination processes ensure the quality, credibility, and usefulness of survey outcomes. These processes enable objective evaluation of data, informed interpretation, and evidence-based decision-making, ultimately contributing to a deeper and more accurate understanding of working conditions in Europe. The thorough documentation of processes and of the different measures to control and assess the quality of its processes and outputs, sets the EWCTS 2021 above most comparative surveys, and at the level of other European and international surveys.

3.2.1 Analysis of quality indicators on reporting and dissemination

This section aims to assess the quality of the reporting and dissemination outcomes of the survey. That is, how survey outputs are documented and made available to the public by Eurofound through different channels. The quality indicators analysed in this section are the following:

Table 22. Quality indicators on reporting and documentation

Number	Criteria	Indicator	Target	Assessment
162	Accessibility	Comprehensive plans and detailed timetables for the various stages/tasks (e.g., cognitive test, sampling, etc are provided)	Y	Target met
163	Accessibility	Comprehensive reports following agreed formats are provided for each stage/task of the process.	Y	Target met
164	Accessibility	Comprehensive documentation on the survey is made available to the public.	Y	Target met

Source: Author's own elaboration, based on EWCS CATI 2021 QAP - Final Version - 25 February 2022

All indicators included in the QAP related to reporting were successfully fulfilled. Thorough plans and schedules for the different phases/activities were provided. The project timeline was consistently revised (Eurofound, 2021b). Both in the Quality Assurance Control Plan and in the individual reports of each process, the timing, and deadlines during which each process was carried out are detailed. Other communications and deadlines were followed via email and were hence not accessible for the assessment. Some delays were incurred at various stages of the project, particularly the fieldwork, as is to be expected in such a complex project, organised for a long period and re-organised for the CATI in a short time, but in general they were well handled by both the contractor and Eurofound with regular updates and controls and did not threaten the overall quality of outputs.

According to Ipsos comments, most of the comprehensive reports followed an agreed format and were provided by for all stages of the survey process as stipulated in the terms or reference. This ensured transparency of the EWCTS 2021 through documentation of the project from preparation to implementation.

Finally, the last accessibility indicator which refers to the documentation of processes and their public availability was also successfully met. Although not included, punctuality should be ensured for the publication of the survey data and its associated documentation at agreed-upon times as it enables interested users, such as researchers, policymakers, and the public, to access information in a timely and transparent manner. This fosters trust in the data and the institution providing it.

While there are only three indicators directly related to reporting several other in the QAP are devoted to the thorough documentation of all processes recognizing its paramount importance not only for accessibility but also accuracy. Such is the case for all processes along the survey's life cycle, as an example solely for the case of weighting: indicator 39 "percentage of countries for which the weighting strategy and procedure are made completely transparent in the weighting report"; indicator 46 "procedure for constructing design weights outlined in sampling report"; indicator 50 "procedure for constructing post-stratification weights outlined in weighting report"; indicator 54 "weight trimming follows the weighting strategy and is fully documented and replicable", etc.

3.2.2 Comparability in reporting and dissemination to gold standards

This section provides a comparative assessment with the European Social Survey (ESS), American Association for Public Opinion Research (AAPOR), Market Research Association (ESOMAR), American Statistical Association (ASA), and the International Social Survey Programme (ISSP), regarding the dissemination process.

The ESS recommends to emphasizes transparency and openness in the dissemination process. The aim is to provide clear documentation of the survey methodology, sampling, and data collection procedures in their reports. The ESS team focuses on providing accessible and comprehensive information about the survey's design, implementation, and results. The ESS contributes to the credibility of survey findings by promoting transparency and ensuring that users can understand the context and limitations of the data.

AAPOR emphasizes accurate and transparent reporting in survey research. Their guidelines stress the importance of documenting survey methods, sample design, and data collection procedures. AAPOR encourages researchers to provide context and explanations for potential sources of bias or limitations in survey results. The goal is to enable readers to understand the survey's methodology and make informed interpretations of the findings.

Another standard is ESOMAR. ESOMAR's principles promote ethical and transparent communication of research findings. Their guidelines emphasise providing comprehensive explanations of research methods, ensuring that findings are presented accurately. ESOMAR encourages researchers to be transparent about data sources, collection methods, and any adjustments or weighting procedures applied to the data. Clear and transparent reporting is essential for maintaining the credibility of research results.

The ASA also emphasizes accurate and appropriate communication of statistical information. Their principles highlight the importance of representing uncertainty associated with estimates and avoiding misleading interpretations. ASA encourages researchers to provide clear explanations of statistical methods and assumptions used in analysis.

ISSP promotes transparent reporting and sharing of survey results. Their guidelines emphasize comprehensive documentation of survey methods and procedures. ISSP encourages researchers to provide detailed explanations of question wording, response options, and any deviations from standard procedures. Transparent reporting ensures that other researchers and users can evaluate and understand the survey process and findings.

According to the AAPOR/WAPOR Task Force Report on Quality in Comparative Surveys (2021), documentation is compulsory for monitoring survey quality, as it is necessary for measuring and comparing '3MC' surveys. It also contributes to determining critical quality dimensions from the survey. Hence, documentation requires data storage, which includes integrating data across time and sources, sustainability, and dissemination. However, according to the report, high-quality documentation is rare in the '3MC' context due to operational difficulties with survey documentation, unequal resources, and research infrastructures (such as within country teams), among other factors. Consequently, '3MC' studies do not present a uniform criterion concerning documentation.

There are diverse options for collecting data:

- Study-level metadata characterises the survey project as a whole and provides metadata for each unit or group of units (such as national or subnational surveys) that comprise a '3MC' project.
- File-level metadata in which the primary components are the file's technical specifications, such as its size, number of variables, number of cases, and enhancements like a checksum to confirm the validity of the original data file if corrupted.
- Variable-level metadata, usually documented by a codebook, and administrative and structural metadata and paradata, which are auxiliary data collections in a survey that describe the process of survey production.

For specific characteristics of data collection, Kallas and Linardis (2010) present several data archives produced in recent years for accumulating, documenting, and disseminating data. Although this process is usually problematic when collecting data from cross-cultural studies in the same data archive (or data metadata repository). Therefore, following specific criteria to homogenise the documentation process is binding. It is basic to follow commonly agreed concepts, measurement patterns, questions, and universes, and it ought to be unified by the coordinating institution in the common agreed language (usually English). Afterwards, the primary entities used in the models are the study, wave, wave instance, source data element and universe-specific data elements: source questionnaire and universe-specific questionnaire, harmonised data file and universe-specific data file, harmonised variable, and universe-specific variables, and lastly, transformations of the universe-specific data element to source data element (Kallas and Linardis, 2010). The main idea is to provide the general information of the study, such as the name, the time and universe instances (if it is part of an established trend) and then harmonise all the information, variables, measurement, and category, and make it available to everyone.

All things considered, EWCTS 2021 and EWCS follow and comply with the criteria presented by the institutions mentioned above. The thorough documentation and transparency in its processes and of the different measures to control and assess the quality of its processes and outputs, sets the EWCTS 2021 at the level of best current standards. Moreover, it is noteworthy its efforts to provide complete information regarding the metadata both at aggregated level and country level with all the considerations required for a better understanding of the data and the survey design process.

3.2.3 Reporting and dissemination quality assessment

This section also focuses on assessing the quality of the survey's output, particularly in terms of reporting and dissemination efforts aimed at making the EWCTS 2021 results findable, accessible, interoperable, and replicable (FAIR). Eurofound's efforts to document its processes and decisions are commendable, and a vast array of reports documenting the different processes and quality assurances and assessments of the survey are produced and made publicly available. There are some details, however, that could improve the accessibility of key documents on the website. Certain inconsistencies can be observed in the presentation of the information, both in terms of the nomenclature of the files and their typology, if we look at the different editions present both on the UK data archive platform and on Eurofound's website. This fact, although apparently trivial, poses inconveniences for users since they have to adapt to different nomenclatures to identify the version of interest. If each edition is modified, it makes it difficult to quickly access the codebook or the translation or paradata document of interest. Additionally, permanent links should be used, or redirections ensured when links become obsolete, since sometimes hyperlinks in the reports are corrupted or no longer work when the documents are updated, or the location changes (E.g. “European Working Conditions Telephone Survey 2021: Technical report” p. 95). As previously discussed, the names and labels could be harmonised, although their content and characteristics might be evident for the people involved in the EWCS, they can be sometimes confusing for the external visitor (E.g. different documents named glossary, or the lack of the year of the survey of reference in the title of a report (E.g. “Working conditions European Working Conditions Survey - Cognitive pretest report”, or “Sixth European Working Conditions Survey – Overview report”, or “European Working Conditions Survey (EWCS) Cognitive Interview Report”).

Eurofound provides public access not only to reports presenting the results of analysis of the data and process/methodology documentation but also to the data and paradata, even at a highly detailed level upon request. Enhancing the connection between all Eurofound documents related to EWCTS 2021 stored at Eurofound’s website and the dataset stored at the UK Data Archive would be beneficial. Accessing the data from outside the UK has a multi-layered process, and it could be worth considering including the dataset in European repository due to the European nature of the data. Nevertheless, the quality and accessibility of the data remain unaffected by the repository of choice. It should be indicated that EWCS and EWCTS 2021 comply with all international standard quality criteria by hosting the data set in a public data repository that guarantees the anonymity of the respondents and stores a protocol for the protection and safeguarding of information, as well as quality standards regarding accessibility, disclosure and transparency of the information contained in its databases.

As discussed, Eurofound maintains a dual strategy of publishing part of the information related to its dataset and the information essential for downloading on its website and using the database in the repository mentioned above. Although this does not represent a problem and is a standard practice in other surveys of similar range both in Europe (European Value Studies, European Parliament Elections) and other international surveys, the presentation and provision of information and documentation presents some difficulties in terms of the user experience that could be easily improved in the future. For example, the microdata is not available under the 'Data' tab of the website, which is devoted to the data explorer, neither is available through the search function. The only way to access it is through EWCS web page only accessible through the 'Surveys' tab. The general search tool prioritizes 'data items', rather than the 'webpages' for EWCS or its associated 'publications'. The 'publications' tab search engine only list one of the many reports when "EWCTS 2021" is searched and a publication on European Works Councils when "EWCS". These issues are all easily solvable and will improve the accessibility of the data and the prolific and high-quality documentation produced.

While the Reporting process does have its own set of analysis indicators and as stated is further considered along other processes indicators, no indicators have been designed to assure a quality dissemination process for the EWCTS 2021. This is reasonable given that dissemination is not directly related with the quality of the processes or outputs, but as previously discussed, the accessibility and use by the public, scientific community and policymakers is in line with the survey objectives and could positively impact the quality by facilitating further analyses carried by other researchers and methodologists about unexplored quality aspects.

The review of both Eurofound and UK Data Archive analytics show a quality dissemination of the survey outputs, both regarding the associated reports, and the data itself. Regarding the EWCTS 2021 dataset, from December 2022 to October 2023, 588 individuals downloaded the file, of which 530 were from European countries (407 without the United Kingdom), accounting for 90% of the total downloads. The countries with the highest downloads were the United Kingdom (124 - 21% of the total), Spain (83 - 14% of the total), and Italy (57 - 9.7% of the total). Of those who downloaded the EWCTS 2021, 24% were affiliated with Postgraduate Universities (142), followed by personnel from Institutes or Higher education institutions with 19% (114), and in third place, personnel from institutes or higher education with 16% (98); 60 were from NGOs or non-profit organizations and 39 from central or local government. 93% of the downloads were for non-commercial purposes (551), while the remaining 7% were for educational purposes (37). 24% of downloads were from people from the academic field of Economics (141), Sociology (96), and in third place, with 36 downloads (6.1%), Management studies and psychology.

The Report "Working conditions in the time of COVID-19: Implications for the future" (2022) was downloaded approximately 3600 times, with the majority 39% (1400) of the retrievals occurring during the first quarter of 2023, followed by the second quarter of the same year with 25% (900), and the last quarter of 2022 with 22% (800). Similar to the report, the first two quarters of 2023 had the highest retrievals. Spain is at the top of the list with 655 consultations (the quality assessment may have affected this data), followed by Ireland with 615 (Eurofound may have affected this data) and France with 511. Sites associated with the EWCTS 2021, including pages related to methodology, questionnaire, etc., had a total of 31,000 visits, the first quarter of 2023 had the highest visitor traffic at 27% (8764), followed by the second quarter of 2023 with 24% (7847), and the fourth quarter of 2022 with 19% (6115).

This data could be further exploited to improve the dissemination process and EWCS recognition and use, facilitating the accessibility and monitoring somehow the fitness for intended use of the survey outputs. The number of downloads and visits can be accessed by Eurofound both for their website and reports and from the UK Data Archive and was also facilitated for this assessment as detailed above, part of the information such as the number of downloads could be made accessible to the public. A section of the website could be devoted to scientific reports and papers which make use of the EWCTS 2021 data similarly to the UK Data Archive under ‘resources’. Mentions in policies, and policy consultations could be disseminated too. The extension of the EWCS to more countries in each edition could be considered a clear indicator of its dissemination success and interest for stakeholders.

4. Conclusions

The transition from the EWCS 2020 CAPI survey to the EWCTS 2021 CATI survey must be considered as an example of rigour, professionalism, and determination to do the best possible work in perhaps the worst conditions. Especially for carrying out a survey in which, due to its idiosyncrasy and its object of study, the execution of personal interviews seems essential.

However, despite the unexpected appearance of COVID-19 and its subsequent effects on society in general and for the development of the survey in particular, Eurofound converted this challenge as an opportunity to extract different learnings: An opportunity to delve deeper into the survey topic and provide real-time information on the effects of a contingency, such as the pandemic, on the working conditions of European citizens. This challenge gives it the originality and almost the exclusivity being able to study the different measures and perceptions in this regard, as well as the effects on the health and well-being of citizens, making it unique in terms of comparability and analytical richness.

On the other hand, Eurofound faced challenges involving:

- deciding to move forward despite the conditions.
- adapting the questionnaire making difficult decisions such as the modularisation of part of the questionnaire due to said transition.
- the transition from CAPI to CATI with all its implications.

However, these challenges resulted in outcomes of maximum interest to any survey methodologist due to the originality and exclusivity of its results.

Overall, the evaluation of the quality of the survey processes is deemed as having a high level of compliance with Eurofound's QAP indicators across all stages. The QAP has served as a relevant, robust, and comprehensive tool to track and control the quality of all processes, along the survey lifecycle from sampling, questionnaire design and development, fieldwork, and weighting. This is particularly notable given the pressing circumstances and the change in administration mode. The quick adaptation of the QAP to serve the purpose of a telephone survey was swift and granted a quality process along the survey lifecycle. There were some minor deviations, but this non-compliance or almost-compliance in some cases, is assessed as having a minimal effect on data quality.

The QAP is generally assessed as a great framework against which to monitor the quality of the survey processes. Although some suggestions have been made on alternative indicators, indicators already in place are generally relevant, appropriate, and comprehensive. If anything, the list could be optimized to facilitate the work of the fieldwork contractor, signalling those that are more relevant for the overall quality of the outputs.

All in all, the processes carried out in the EWCTS 2021 is (in most cases) up to best practices and standards on '3MC' surveys.

The questionnaire development process incorporated many current best practices such as: consultation with subject matter experts and stakeholders, overall assessment by an expert in '3MC' survey methodology, translation following a simplified TRAPD approach (translation, review, adjudication, pretest, documentation), harmonisation and adaptation, and a sizable investment in training and piloting pretesting. This questionnaire made use of the advance translation, cognitive test and full TRAPD translation previously produced for the EWCS 2020 since the questionnaire was only slightly modified.

The questionnaire's adaptation to suit a telephone interview involved its shortening, reducing response scales, and adapting and reducing the introduction and final questions via a planned missingness design with a modularised approach. The experimental nature of this decision has been the precursor of much debate in the academic environment about its effects and convenience. So much so that the European Social Survey itself has included this mechanism in its latest round, and other European and international surveys are debating including these modalities. Although these processes require exhaustive reviews and control mechanisms around representativeness, non-response, sampling, and selection bias, etc., as has been seen in this evaluation report, this does not detract from the effort made and the experience acquired. Although this process has ensured the continuation of most trend questions allowing for valuable data comparisons or tracking, the representativeness and comparability has been evidently affected and comparisons should be made with due caution.

The inclusion of cognitive testing (in the original EWCS 2020) process demonstrates Eurofound's commitment to ensuring that survey questions are not only clear but also interpreted consistently by respondents. This level of rigour is up to best current standards and is vital for comparable and functionally equivalent data collection. Although the data collection mode should be considered in cognitive testing, the decision to make use of previously carried processes is reasonable given the circumstances, time, and budget constraints. Some recommendations are provided in the next section regarding the standardization of methodologies, the possible inclusion of web probing, and additional sample targets.

Eurofound's ability to manage translations in a large number of languages is a testament to its commitment to inclusivity and accessibility as well as accuracy and comparability. This effort ensures that respondents across Europe could participate in the survey comfortably, contributing to both response rate and comparability, ensuring the functional equivalence of questions across languages and cultures. This effort included the advance translation of the questionnaire (for EWCS 2020) and the full TRAPD translation for EWCS 2020 original questionnaire and simplified approach for EWCTS 2021. Overall, the processes retained their quality and included current best standards of '3MC' questionnaire design, some recommendations are made in the next section, particularly to ensure a team approach to reviewing and adjudication in translations and solid cognitive test standards.

Despite the need for different sample designs due to changing conditions, the survey has successfully adjusted its sample calibration and treatment procedures accordingly. Overall, the quality assessment of the weighting concludes that the EWCTS 2021 has followed sound principles for its sampling design and weighting procedure. All participating countries implemented a probability-based sample design, using a high-quality sampling frame, and developed sampling strategies with the objective of minimising sampling errors and maximising efficiency. Two different sampling designs were used: a simple random sample by Random Digit Dialling (RDD) in all countries but Sweden where stratified sample with proportional allocation was selected from the sampling.

According to Kish (1994), sample collecting process and design may adapt to each national resources and its potential to account for increasing probabilities of gathering all population elements. Therefore, the flexibility showed by Eurofound to adapt sampling designs (such as the case of Sweden) is a demonstration of flexibility required for this kind of surveys as being done in other Gold Standards (ESS, American National Election Studies, World Value Survey, European Value Studies).

The sample was carefully designed to be comparable across all the participant countries, the sampling frames were relatively wide regarding coverage, and the sample sizes large enough to produce reliable national estimates. The estimated coverage in some countries is below 90%, which can result in coverage biases since the use of mobile phones varies considerably by sociodemographic characteristics.

The possibility of combining RDD and face-to-face samples was unfeasible due to the COVID-19 restrictions that were in place in the vast majority of Europe. There was also the possibility of combining RDD and landline samples, but due to the increasingly low coverage of landlines, that would have not resulted in a noticeable increase in coverage. Determinations of sample size and necessary precision are key issues relating to sample design in '3MC' surveys (AAPOR, 2021), and therefore some details from the procedure to allocate sample size should be included in the sampling report.

The fieldwork process was meticulously planned and closely monitored, allowing for the prompt detection and resolution of issues, resulting in a high-quality sample. Special mention to Ipsos, its fieldwork contractor that worked hand in hand with Eurofound along the process. The contractor demonstrated adaptability by transitioning to CATI and incorporating random modules of completion. This approach allowed for a proportional distribution of completed surveys across modules, showcasing the organization's ability to evolve in its data collection methods. Although some issues arose in the implementation of the planned missingness design, Ipsos effectively addressed the issue involving a secondary check of completed quotas.

The weighting system was implemented following regular standards used in calibration, with a proactive construction of design weights taking over coverage into account, a calibration procedure in various steps to avoid further problems, using auxiliary variables that may have correlations with potential variables of interest, and using linear bounded distances which avoids further weight trimming. In addition, the analysis of the weighting procedure was thoroughly documented in the Sampling & Weighting Report, including the extent of the adjustment that had to be done in the calibration step, a comparison between unweighted and weighted estimates from the survey and estimates from the LFS of each country in a number of monitoring variables (allowing for a bias analysis), and an analysis of the design effect in each country.

Weighting adjustments were applied to minimize nonresponse bias, a common challenge in telephone surveys, by utilizing available auxiliary variables. Nonresponse has become an important issue in sample surveys, especially with declining participation rates, particularly evident in telephone surveys (Beullens et al., 2018). This survey is no exception, and a notable problem is the very high nonresponse rate, a common issue in CATI surveys also experienced by other surveys carried during the pandemic such as the Labour Force Survey or Americas Barometer (Hox & De Leeuw, 1994; De Rada, 2015, AAPOR, 2021; Castorena et al., 2022; Eurostat, 2022). It is important to note that the nonresponse rate alone does not directly indicate the non-response bias for a specific survey and in this survey the percentage of non-response is quite balanced in the categories of the sociodemographic variables considered, so the reweighting methods used were useful to reduce the observed bias. There are other approaches for handling unit nonresponse, such as the propensity-score weighting (Sverchkov, 2008, Riddles et al. 2016), which increases the sampling weights of the respondents using their inverse response probabilities. Unfortunately, it is difficult to apply these techniques in this survey because no data is available about non-respondents.

Generally, the use of a standardised CATI instrument in the EWCTS 2021 facilitated a standardised collection of paradata. In relation to the microdata, the quality assessment and comparison to previous rounds was difficult given the change in the mode of administration, the questionnaire adaptation to interviews conducted over the phone instead of face-to-face, which entailed the use of different (reduced) response scales, and inclusion, and exclusion of questions. In addition, changes in the results could be a result of real changes in working conditions during COVID-19. In terms of survey quality, there were no relevant issues found that could have seriously compromised the internal and external validity, reliability, or overall quality of the survey results.

Furthermore, the contractor showed openness to improvement by considering recommendations from local agencies on adapting materials and training procedures. This willingness to evolve demonstrates a commitment to enhancing data collection processes. Ipsos addressed the importance of data quality, particularly when dealing with variables like working hours, where standardization and differences between countries can affect reliability. Their attention to these details underscores their commitment to accurate data.

Another aspect to highlight is the execution times and the quality criteria maintained over time through all the different processes. Although in some cases, such as the delivery of documentation or the training of interviewers and in the decision-making processes, some delays were found, these have not had any major effect on the final result, being perfectly understandable given the circumstances and the uncertainty scenario of that moment on all personal, professional and technical levels.

Despite the difficulties expressed throughout this report for the execution of the survey, no major issues have been detected with a severe impact on quality, which is very notable if we take into account that they have gone through a triple quality evaluation process: its own, that carried out by the company hired to manage the field and by this one.

Overall, for this survey edition all countries implemented a probability-based sample design, using a high-quality sampling frame, and developed sampling strategies with the objective of minimising sampling errors and therefore maximising efficiency despite the inconveniences caused by the COVID-19. Although the conditions imposed the use of different sample designs the sample calibration and treatment procedures have been adapted to these new sample designs. The deployment of the fieldwork has been carefully undertaken and monitored from start to finish, which has allowed the detection of issues and fix them immediately, leading to a high-quality sample. The weighting adjustments ensured that the nonresponse bias, which is very common in telephone surveys, is reduced to a minimum using the available information on auxiliary variables. Regarding external validity, the comparable variables estimates are quite similar to those from the LFS. A detailed study would have to be carried out using the LFS microdata to reach definitive conclusions.

Finally, this report would not like to miss out on highlighting Eurofound’s concern for transparency and good governance demonstrated by the very large array of information made available, not only for the correct use of the survey and its data set, but also to understand the entire procedure. Such an amount of documentation requires, in turn, a notable effort to maintain the workflow, the nomenclatures and information on presentation formats, as well as a map, website or a guide that facilitates the user experience. In this sense, an exercise of including indicators or holding a focus group to analyse the navigability and usability of its resources could help to a large extent to mitigate the difficulties posed by hosting the documentation almost in parallel in two different platforms: the UK Data Archive, which guarantees the international standards of accessibility and usability of the database as well anonymisation and preservation of the data sets, and Eurofound, the other platform which hosts the rest of the documentation on its website (methodological reports, publications, data explorer, etc.).

Ultimately, this report concludes that EWCTS 2021 meets all the quality standards of international surveys, presenting an opportunity for researchers, companies and institutions that want to learn and make future decisions about working conditions in Europe, their effect on the state of individual and collective well-being and especially the effects of COVID-19, teleworking, and job uncertainty on the working conditions of European citizens.

5. Recommendations

Finally, one of the most important outcomes of this Quality Assessment Report are the recommendations for maintaining and increasing the quality and hence the relevance of the EWCS. To offer the most feasible recommendations, the methodology proposed implies a convergence process (Thomas et al., 2021; Van Praag, 2021) by which critical actors (interviewees, along with the evaluation research team) propose and prioritise the recommendations. Therefore, the highest ranked are those more relevant in terms of efficiency, feasibility, and impact in the quality of the outcome, providing a handy list for Eurofound.

5.1 Quality assurance recommendations

Reduce or prioritise the Quality Assurance Plan

1. The QAP is generally assessed as a good framework against which to monitor the quality of the survey processes. The plan could however be reduced, or specific elements prioritised to facilitate the work of the fieldwork contractor, signalling those indicators more relevant for the overall quality of the outputs, which generally but not exclusively are those related to Accuracy. For the EWCTS 2021 an already reduced list of 134 indicators were considered, 25 directly dependent on Eurofound, and 105 on the contractor, with 4 indicators shared along the parts.

Ensure a consistent labelling and workflow

2. Eurofound's team develops both, internal and publicly, a large amount of documentation which enrich and ensure the quality of the work done throughout each EWCS edition, even under critical circumstances such as the COVID-19. However, this effort does not always pay-off in terms of clarity and accessibility for the interested user. CESSDA (Consortium of European Social Science Data Archives) and ERIC (European Research Infrastructure Consortium) present a list of suggestions for naming and storing properly the back-office or internal documentation of any research process and how to allocate properly for its dissemination purpose. Keeping a consistent and repeated structure of the technical reports, metadata information, questionnaires, survey results and its potential updates or revisions throughout time, would facilitate internal work but, more important, the overall impact of the survey.

Expand information on the decision making in the reports

3. Although the QAP indicators and different processes logs show a high compliance with different quality assurance steps and their documentation, sometimes information on decisions on final outcomes is missing. For example, decisions on questions added, modified, or eliminated in relation to cognitive tests or pilot results are not well documented.

5.2 Questionnaire planning and design recommendations

5.2.1 Questionnaire design

Develop an analysis plan to reduce the questionnaire

1. Continue working to reduce the survey length where possible as it is one of the main reasons for refusals or drop-outs and a priority given the increase of non-response in all survey modes (Hox & De Leeuw, 1994; Galesic & Bosnjak, 2009). Respondent engagement should be considered, and the design and length of the questionnaire adapted to ensure greater participation and the accuracy of the data collected (Groves et al., 2009). Although this is something common in comparative surveys and there is always a trade-off in terms of quality, in which an improvement in the response rate, engagement of the participants and accuracy would be made at the cost of the relevance of the survey, continue considering accuracy and comparison error by retaining essential survey questions and control variables and reducing the questionnaire where possible. This could be addressed by continuing the initiated work on defining the rationale for including questions and further developing how they have been or should be exploited in an Analysis Plan or further developing the Glossary "measurement objectives", which on the long run could contribute to the prioritization of questions. Another option is to continue testing planned missingness design or the split or modularization of the questionnaire; as well as testing the duration of the survey across profiles and paths to find where efforts should be increased. Focus groups with groups of respondents with a specific less typical profile such as those conducted in the ESS could be useful in finding the right formulation for questions.

Continue and develop the glossary and concordance grid.

2. Providing the Glossary to the translators helped to support the functionally equivalent translation of the key terms, while the Concordance Grid serves as an aid to compare different EWCS questionnaires and making these available is considered good practice that build on recommendations from previous assessments. The recommendation is to work on refining these documents to ensure their timely update and availability. Considerations from the glossary, like the rationale for including or modifying questions, the expert assessment and the source or international standard from which the question was taken or adapted could enhance not only the accuracy but the accessibility and comparability of the survey, so it is recommended to continue this practice and to make the Glossary fully or partially available to the public, either by itself or by including some information into the Concordance Grid.

Ensure specific demographics remain engaged.

3. Ensure specific demographics of interest for working conditions can be and remain engaged, especially certain groups with lower educational levels and older workers which had difficulties understanding certain terms and questions or needed more time to answer the survey, particularly those questions that were originally designed to be accompanied by show cards, as mentioned in the Pilot Report. A way to address this issue would be to specifically analyse item nonresponse in the pilot in relation to these variables (age and level of education) in the cognitive tests sample design

Implement Methodological Workshops.

4. While stakeholders are indeed involved at several stages of the questionnaire design, consideration should be given to survey methodologists and those who regularly exploit the survey data. To this end Stakeholders Methodological Workshops similar to those employed by the ESS, WVS, EVS or PISA could be implemented every couple of years. The purpose should be to generate feedback with subject matter experts to potentially improve the survey design. This involves providing a more detailed explanation of why certain questions are retained in the questionnaire, why others are eliminated, or how questions are distributed across different modules.

5.2.2 Cognitive testing

Set minimum standards for the methodology and reporting of cognitive tests.

1. The use of different pretesting strategies from expert review to cognitive testing, and piloting ensures that the EWCS complies with '3MC' best current standards. However, the methodology and form of reporting the cognitive tests results differs widely across waves in terms of expertise of the interviewers, the techniques employed and type of probing, and the way in which the findings and recommendations are presented with *verbatim* or direct quotes or a ranking system. This could point to an area for improvement in the QAP, which notwithstanding the methodology used and allowing Eurofound to adapt to the circumstances of time and budgetary constraints, could set standards on how the test should be carried and reported. The reports should in any case include information justifying the selection of questions and countries in which the cognitive tests were carried (Collins, 2003; Willis, 2004).

Adapt cognitive test to new administration modes.

2. We strongly recommend to always adapt cognitive tests to the mode of administration. The recommendation is to always test at least the screening questions, new questions, ISCO and NACE questions, and those few questions that have been repeatedly deemed difficult even if assisted by a person and showcards. That is particularly important when new administration modes are to be applied, whether CATI or CAWI.

Consider web probing.

3. If the budget allows, we recommend reinstating web probing as already carried out in the post test of the 6th EWCS. While cognitive interviewing would serve in-depth exploration of new or problematic questions, particularly in populations difficult to reach online, web probing can offer some insights on their prevalence, validate questions' constructs, and potentially allow for the extension of the exercise to other countries with additional language families. This is particularly relevant if the next wave is carried as a push to web CAWI given lack of supervision and assistance in the interview.

Consider web probing in the main survey.

1. The CAWI will permit the pretesting of different question wordings and cues, it also offers the chance to implement probing in the main data collection for selected questions deemed problematic. This offers the quite exceptional chance to web probe the validity and equivalence of questions in a probabilistic '3MC' sample, and to compare in-person cognitive test and web probes across cultures (Behr et al., 2014; Metinger, 2017). As this could, however, have an adverse impact on the accuracy of the data by increasing item non-response, which is already particularly high in online surveys, survey breaks, or shifts in response behavior. An alternative would be to apply it to only to a subsample to control these

effects (Schuman, 1966; Behr et al., 2017) or apply closed-ended probes (Scanlon, 2019) designed from the cognitive interviews.

Include additional variables in the cognitive test sample.

2. As stated, including other relevant variables, at least level of education and age (such as including an elderly group), in the cognitive tests sample is recommended.

5.2.3 Translation

Ensure a team approach in the review and adjudication.

1. The quality of translations would benefit from a direct exchange even if conducted through a video conference. The review and adjudication process resembled more of a stepwise consecutive review rather than a direct exchange, although it must be noted it happened at a time in which physical presence was not possible. The simplified TRAPD approach used in the EWCTS 2021 was justified given the circumstances and in line with best standards. Furthermore, it was developed from the previously fully TRAPD translated questionnaire for the CAPI 2020 with only minor changes. The review process however took place largely in writing through the Translation Template, where the adjudicator described the issues encountered. The file shows that corrections were done and argued in detail when needed, but discussions if any, took place via email. Although there were online meetings for the harmonisation process albeit only for German and French, it is our assessment that a personal meeting even if online would have been a better approach, more respectful of the team approach to TRAPD translation, which would have ensured the quality and the minimisation of measurement and equivalence errors to a higher standard at little or no extra cost.

Ensure the pretesting of all languages

2. Although the extent of changes from the pilot is limited and generally refers to issues common to all language, which would certainly be detected in a sizable pilot, some languages, particularly minoritarian languages, could go un-pretested both at the cognitive test (reasonably) and pilot, since the latter follows a random probability approach. A minimum quota could be set for all languages to ensure that they are all covered in case some are not captured in the random allocation, to ensure a the functional equivalence of all translations.

5.3 Fieldwork recommendations

Reduce language barriers

1. Language is paramount to avoid nonresponse error in telephone and web surveys. CAPI techniques can help with refusals, but translating the questionnaire into various languages, including official, regional, or immigrant languages, particularly if they represent a big proportion of the population can further boost response rates. This is especially important for push-to-web surveys with less direct interaction. In this sense, it should be noted that the EWCTS already increased notably the number of languages from previous editions. Mixed modes or CAWI offer the possibility to answer the questionnaire in any of the already scripted

languages regardless of the country, which would include many immigrant populations both intra and extra European.

Develop more visual or interactive training materials

2. Occupation and sector of activity questions which use the ISCO and NACE international standard classification are vital but capturing this information to a four-digit level is very demanding for the interviewer and interviewees as shown in all the cognitive and pilot testing. As is the screening question or eligibility criteria. It is important to note that the persistence of issues in certain questions is deemed to be related not to a poor questionnaire design or insufficient pretesting, but rather to the complex nature of the studied topic. The approach used where interviewees explain in their own words - following probes by the interviewer - their activity and sector is an intelligent, well established and effective solution, however it is very dependent on the interviewer (which can introduce a bias although measures are taken to reduce it, for example limiting the number of interviews per interviewer) and it is a difficult path to follow if a CAWI system is to be used in the future. On the interviewer side, the train-the-trainer approach (TtT), and materials used are up to best current practices. However, the recurrent character of the survey and its dependency on those questions could merit an investment in newer approaches like developing more visual information, or short videos or animations with updatable subtitles if needed, and especially online exercises or tests on how to probe, directly addressed at interviewers. At the bare minimum, it is recommended to follow fieldwork contractor suggestions on fieldwork materials which suggest reducing explanations about the history of the survey and to focus more on the above questions.

Exclude willingness to be recontacted from interview time calculations

3. Given that a proportion of the extended interview length is due to questions regarding the willingness to be recontacted for follow-up research, it may be worth considering excluding these questions from the survey length calculation. Doing so would provide a more accurate estimate of the actual length of the survey, focusing only on the questions relevant to the data collection of the study in question.

Control implementation of planned missingness designs

4. If a planned missingness design (modularisation or split questionnaire designs) is to be used again, adequate measures or indicators to follow its execution by the fieldwork contractor need to be implemented in order to minimise issues with reverse scales or sampling allocation.

5.4 Sampling & weighting recommendations

Implementing new variance estimators

1. Look for census data and administrative registers or using methods to calibrate to estimated control totals rather than to population values. In this case it may be needed to account for the variance of the estimated control totals to ensure that calibrated estimates appropriately reflect sampling error of both the primary survey of interest and the survey from which the control totals were estimated. Dever and Valliant (2010) develop and evaluate variance estimators for point estimates with weights that contain a

poststratification adjustment to a set of survey estimated control totals, which is the case of this survey.

Using multiple frame designs in some countries

2. Frames can and will vary in a '3MC' survey, and this variability in and of itself does not necessarily challenge data comparability (AAPOR, 2021). However, the quality of available frames in this study differs across countries with regards to coverage, leading to significant differences in degree of population representation. For this reason, the use of multiple frame sampling in those countries where coverage has been low is recommended wherever possible, combining landline, mobile and online surveys (Arcos et al. 2015, Metcalf and Scott 2009, Mecatti and Singh 2014)

Selecting variables for calibration by country

3. Another recommendation would be to separate weighting procedure for each country selecting the best variables for calibration, either because they explain non-response better than other variables that may fit in other countries, or because the reference statistics for them are available (or its estimations are more accurate) (Beaumont 2008, Haziza and Beaumont 2017).

Include details from the procedure to allocate sample size in the sampling report

4. Determinations of sample size and necessary precision are key issues relating to sample design in '3MC' surveys (AAPOR, 2021), and therefore some details from the procedure to allocate sample size should be included in the sampling report.

Leave adjustment uncapped if CATI is used again

5. This evaluation considers that the impact of leaving the number of phones uncapped should not have an impact on the variance, given that very few people may have three or more mobile phones, and in that case the weights for those people would be lower, meaning that the impact on the final estimates would be limited as well. Capping the number of phones could, however, have an impact on the bias because the population with more mobile phones is different, according to the words of one of the interviewees. This is a relevant fact, as the variance can be treated afterwards (for example, by bounding weights) but the bias induced by incorrectly weighting individuals is harder to treat. On the other hand, the number of phones was self-reported by the respondents, meaning that some kind of measurement error could be taking place in this variable as well. Finally, if the adjustment factor must be capped, we agree with the recommendations from one of the interviewees that the capping should depend on the characteristics of each country, rather than imposing a single limit.

5.5 Microdata & paradata recommendations

Include information on the original weights of the design in the data file

1. Since it has been verified that the patterns of non-response follow different behaviours by country, it would be convenient to have information on the original weights of the design in the data file so that any researcher wanting to make use of them can try another weighting with more relevant and extensive information that is available in each country in order to obtain more precise estimates of the parameters.

Develop an advanced response analysis from all participating countries

2. Nonresponse and the resulting nonresponse bias reducing cross-national comparability of survey estimates, remains a cause of concern. Eurofound should consider performing an advanced response analysis for all participating countries. For it, data must be available for all sample units, be collected, and coded consistently across all cases and in all participating countries. In the EWCTS 2021, data collected for non-interview cases was limited to number of contacts, call time, and characteristics of the interviewer. To obtain additional observational data about the person to interview would provide more comprehensive information about non-respondents and how they might differ from respondents. This could give an idea of the potential bias.

Use of paradata to explore non-response bias.

3. Process data could be used to analyse non-response bias. The data file includes a variety of important process data that can be used to optimize the fieldwork and address unit non-response, for example through multilevel models (van de Vijver and Leung, 1997) or machine learning models (Kufner et al 2022). An example of the use of paradata in non-response analysis can be seen in the European Company Survey 2019 (Eurofound, 2019). This analysis will serve to characterize response probability and could help reduce nonresponse to specific questions in future editions of the survey.

5.5 Reporting and dissemination recommendations

Harmonisation between Eurofound´s website and the UK Data Archive.

1. With the aim of enhancing the accessibility of the reporting and data of the EWCTS 2021, we recommend the harmonisation of the documentation available in each wave on both, Eurofound and the UK Data Archive websites. In this regard, the documents provided differ for various editions. This could create a burden for those seeking access to the same documents for comparative or longitudinal studies. In this context, Eurofound could adopt from now onwards a consistent approach for uploading the required documentation for each wave.

Harmonise nomenclatures

2. Unify and standardise the nomenclature style guidelines for reports published both on the UK Data Archive and across different sections of the waves on the Eurofound website, for example always using the year of the survey of reference in a report, to enhance the user experience in identifying documents across editions of the EWCS. That will follow gold standards such as ESS, WVS, EVS, among others.

Use permanent links or redirections

3. Ensure permanent links are used, or redirections are in place when links become obsolete, since sometimes hyperlinks in the reports are corrupted or no longer work when the documents are updated, or the location changes.

Enhance the user´s experience

4. Ensure the accessibility of the data and publications by enhancing accessibility and the user’s experience, introduce navigation tools to locate important sections of the survey and its corresponding documentation, or review queries and keywords. For example, microdata is

not available in the "Data" tab of the website, which is dedicated to the data explorer, nor is it accessible through the search function. Prioritising the most crucial survey sections for statistical analysis could be implemented for future editions. Likewise, a section of the website could be devoted to scientific reports and papers which make use of the data similarly to the UK Data Archive under ‘resources’ which will promote the survey and show the fitness for intended use of the produced data. A succinct usability test from as user external to Eurofound could inform this endeavour.

Update and make public the Concordance Grid and Glossary

5. It is recommended to continue and ensure the timely update of the Concordance Grid that compares different EWCS questionnaires. It could also include information from the Glossary about the rationale for the inclusion of questions. Both are considered good practices, which build on recommendations from previous assessments

Consider new access routes to the UK Data Archive or other data repositories

6. UK Data Service complies with all the quality standards required for an international survey like EWCTS 2021. Furthermore, it also has a common procedure to register and declare the purpose of accessing the data of interest. However, it has some access barriers: the user should belong to a UK institution or affiliated with an accredited educational institution, which is considered an additional burden especially because of the European nature of the data, and the navigation or search for the database is less intuitive than similar archives. Although it should be noted that the repository of choice will not affect the quality of the data itself, and that the UK Data Archive complies with all quality standards, other options could be considered if accessibility is taken into account.

6. References

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

AAPOR (2021), *AAPOR/WAPOR Task Force Report on Quality in Comparative Surveys*.

AAPOR (undated), *Response rates calculator.*

Arcos, A., Rueda, M., Trujillo, M., Molina, D. (2015). Review of Estimation Methods for Landline and Cell Phone Surveys. Sociological Methods & Research, 44(3), 458-485. ￼

Austin, P. C. (2008), 'Critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'*. Statistics in medicine*, 27(12), 2037-2049.

Beaumont, J. F. (2008*), '*A new approach to weighting and inference in sample surveys '. *Biometrika*, 95(3), 539-553.

Behr, D. (2018), *Survey Design and Methodology Cross-cultural survey methods.*

Behr, D., Bandilla, W., Kaczmirek, L., & Braun, M. (2014). Cognitive probes in web surveys: on the effect of different text box size and probing exposure on response quality. Social Science Computer Review, 32(4), 524-533.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing-implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions (Version 1.0).Hadler, P. (2023). The Effects of Open-Ended Probes on Closed Survey Questions in Web Surveys. Sociological Methods & Research, 0(0). https://doi.org/10.1177/00491241231176846

Belisario, J. S. M., Jamsek, J., Huckvale, K., O'Donoghue, J., Morrison, C. P., Car, J. (2015). *Comparison of self-administered survey questionnaire responses collected using mobile apps versus other methods*.

Biemer, 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.

Blumberg, S. J., Luke, J. V. (2021). Wireless substitution: early release of estimates From the National Health Interview Survey, July-December 2020. National Health Interview Survey early release program, National Center for Health Statistics. https://dx.doi.org/10.15620/cdc:108678

Brancato, G., Macchia, S., Murgia, M., Signore, M., Simeoni, G., Blanke, K., & Hoffmeyer-Zlotnik, J. (2006). Handbook of recommended practices for questionnaire development and testing in the European statistical system. *European Statistical System.*

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., Stürmer, T. (2006). 'Variable selection for propensity score models *'. American journal of epidemiology*, 163(12), 1149-1156.

Campbell, D. T., Fiske, D. W. (1959). 'Convergent and discriminant validation by the multitrait-multimethod matrix *'. Psychological bulletin*, 56(2), 81.

Castorena, O.; Montalvo J.D.; Pizzolitto, G.; Plutowski, L.; Schweizer-Robinson, V., and Wilson, C. J. (2022). Methodological Note #011 Response Rates in LAPOP's 2021 AmericasBarometer. Wilson Vanderbilt University

Chang, T., and P. S. Kott (2008), 'Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model ', *Biometrika*, 95, 555–571.

Chen, K. T. (2016). *Using LASSO to Calibrate Non-probability Samples using Probability Samples* (Doctoral dissertation).

Chen, J. K. T., Valliant, R. L., Elliott, M. R. (2019). ‘Calibrating non-probability surveys to estimated control totals usingLASSO, with an application to political polling ‘. *Journal of the Royal Statistical Society Series C*, 68(3), 657-681.

Christensen, A. I., Lau, C. J., Kristensen, P. L., Johnsen, S. B., Wingstrand, A., et. al, (2022). ‘The Danish National Health Survey: Study design, response rate and respondent characteristics in 2010, 2013 and 2017 ‘. *Scandinavian journal of public health*, 50(2), 180-188.

Clark R.G.; Chambers R.L. (2008) ‘Adaptive calibration for prediction of finite population totals ‘. *Sur Method*; 34: 163–172.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of life research, 12, 229-238.*

Da Silva, D.N.; Opsomer, J.D. (2009) *Properties of the weighting cell estimator under a nonparametric response mechanism*.

Denzin, N. K. (2010). ‘Moments, mixed methods, and paradigm dialogs ‘. *Qualitative inquiry*, 16(6), 419-427.

Desiree, S., Lenaerts, K. (2020). *European Company Survey 2019: Data quality assessment-Methodology Annex*.

Destatis (2022). Equipment of households (Continuous household budget surveys): Germany, reference date, consumer durables, age of the main income earner. Table 63111-0005.

Dever, J.A.; Valliant, R. (2010). *A comparison of variance estimators for poststratification to estimated control totals. Survey Methodology*, 36, 1, 45-56.

Deville, J.C. (2000), ‘Generalized Calibration and Application to Weighting for Non-response ‘, *Proceedings in Computational Statistics*, pp. 65–76.

Deville, J.C., Särndal, C.E. (1992) ‘Calibration Estimators in Survey Sampling ‘. *Journal of the American Statistical Association*, 87, 376–382.

Devins, G. M., Beiser, M., Dion, R., Pelletier, L. G., Edwards, R. G. (1997). ‘Cross-cultural measurements of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translations of the affect balance scale ‘. *American Journal of Public Health*, 87(5), 794–799.

De Rada, V. D., & Portilla, I. (2015). Encuestas telefónicas: estrategias para mejorar la colaboración. Revista Perspectiva Empresarial, 2(1), 97-115.

Ebbs, D., Wry, E. (2016). *Translation and layout verification for PIRLS 2016*. *Methods and procedures in PIRLS*, 1-15.

Edgar, J., Murphy, J., Et Keating, M. (2016). Comparing traditional and crowdsourcing methods for pretesting survey questions. SAGE Open, October-December, 1-14. doi: 10.1177/2158244016671770.

Eurofound (2019)   European Company Survey 2019: Data quality assessment

Eurofound (2021a) *European Working Conditions Telephone Survey 2021*

Eurofound (2021b) *EWCTS 2021: Translation report.*

Eurofound (2021c) *EWCTS 2021: Data validation and editing report*

Eurofound (2021d) *EWCTS 2021: Sampling & Weighting report.*

Eurofound (2022), *Working conditions in the time of COVID-19: Implications for the future, European Working Conditions Telephone Survey 2021 series, Publications Office of the European Union, Luxembourg.*

European Statistical System (2019) *Quality Assurance Framework – 2.0 Version.*

Eurostat (undated) *Eurostat Communication and Dissemination Strategy 2021 – 2024.*

Eurostat (2022). Quality report of the European Union Labour Force Survey 2020.

Ferri-García, R., Rueda, M. D. M. (2018). *Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. SORT: statistics and operations research transactions*, 42(2), 159-182.

Fielding, N., Fielding, J. (1986). *Linking data.* London, England: SAGE.

Flick, U. (2002). *An introduction to qualitative research* (2nd ed.). London: Sage Publications.

Fowler, S., & B. Willis, G. (2020). The practice of cognitive interviewing through web probing. *Advances in questionnaire design, development, evaluation and testing*, 451-469.

Fuller, W.A. (1998). 'Replication variance estimation for two-phase samples *'. Statistica Sinica*, 8: 1153-1164.

Ghirelli, N., Lynn, P., Dorer, B., Schwarz, H., Kappelhof, J., van de Maat, J., Kessler, G., Briceno-Rosas, R., Rød, L-M. (2022). ESS9 Overall Fieldwork and Data Quality Report, GESIS.

Giménez-Nadal, J. I., Molina, J. A., & Velilla, J. (2022). 'Trends in commuting time of European workers: A cross-country analysis '. Transport Policy, 116, 327-342.

Goerman, P., Meyers, M., García Trejo, Y. (2018). 'The place of expert review in translation and questionnaire evaluation for hard-to-count populations in national surveys '. In GESIS Symposium on *Surveying the Migrant Population: Consideration of Linguistic and Cultural Aspects* (Vol. 19, pp. 29-41). DEU.

Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (ed.s) (2001). *Survey Nonresponse*. Chichester: John Wiley & Sons

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer and Roger Tourangeau (2009) *Survey Methodology*. Second Edition, New Jersey: John Wiley & Sons Inc,

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2011). *Survey methodology*. John Wiley & Sons.

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., Peytcheva, E. (2008). 'The impact of nonresponse rates on nonresponse bias: a meta-analysis '. *Public opinion quarterly*, 72(2), 167-189.

Hambleton, R.; Yu, J.; Slater, S. (1999) *Field test of the ITC Guidelines for adapting educational and psychological tests*.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., et al. (Eds.). (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. John Wiley & Sons.

Hartley, H.O. (1962) 'Multiple frame surveys, Proceedings of the American Statistical Association *', Social Statistics Sections*, pp. 203–206, 1962.

Haziza, D., Beaumont, J. F. (2017). 'Construction of Weights in Surveys: A Review *'. Statistical science*, 32(2), 206-226.

Haziza D, Lesage É (2016) *A discussion of weighting procedures for unit nonresponse*. J Off Stat 32(1):129

Hirano, K., Imbens, G. W. (2001). *Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes research methodology*, 2, 259-278.

Hox, J. J., De Leeuw, E. D. (1994). 'A comparison of nonresponse in mail, telephone, and face-to-face surveys: Applying multilevel modelling to meta-analysis '. *Quality and Quantity*, 28(4), 329-344.

Hui, C. H., Triandis, H. C. (1985). 'Measurement in cross-cultural psychology: A review and comparison of strategies '. *Journal of Cross-Cultural Psychology*, 16(2), 131–152.

ILO (International Labour Organization) (10-19 October 2018). *Data collection guidelines for ICSE-18*. Geneva.

INE (2022). ICT product installed in dwellings. Survey on Equipment and Use of Information and Communication Technologies in Households.

International Test Commission (2012). International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores.

Ipsos (2021a) EWCTS 2021 *Quality Assurance and Control Report.*

Ipsos (2021b) *EWCTS 2021 Pilot Report.*

Ipsos (2021c) *EWCTS 2021 Technical Report.*

Ipsos (2021d) *EWCTS 2021 Sampling and Weighting Report.*

Jahoda, M., Lazarsfeld, P., Zeisl, R. (1976). *Marienthal: The sociography of an unemployed community*. London, England: Tavistock

Johnson, T. P., Pennell, B. E., Stoop, I. A., Dorer, B. (Eds.). (2018). *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts ('3MC')*. John Wiley & Sons.

Jowell (2017), *ESS-Impact-study-Final-report*

Kallas, J. Linardis, A. (2010). *'*A Documentation Model for Comparative Research Based on Harmonization Strategies '. *IASSIST Quarterly*. 32. 12. 10.29173/iq652.

Kish, L. (1994), "Multipopulation survey designs: five types with seven shared aspects," International Statistical. Review 62, 167-186.

Kott, P. S. (2005). *"No" Is the Easiest Answer: Using Calibration to Assess Nonignorable Nonresponse in the 2002 Census of Agriculture*.

Kott, P. S. (2006). 'Using calibration weighting to adjust for nonresponse and coverage errors *'. Survey Methodology*, 32(2), 133.

Kott, P. S. (2016). *Calibration weighting in survey sampling*. Wiley Interdisciplinary Reviews: Computational Statistics, 8(1), 39-53.

Kott, P. S., Chang, T. (2010). 'Using calibration weighting to adjust for nonignorable unit nonresponse *'. Journal of the American Statistical Association*, 105(491), 1265-1275.

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

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

Küfner B., Sakshaug J.W., Zins S. (2022). Analysing Establishment Survey Non-Response Using Administrative Data and Machine Learning, Journal of the Royal Statistical Society Series A: Statistics in Society, 185(2), 310–342. https://doi.org/10.1111/rssa.12942

Lee, S. (2006) 'Propensity score adjustment as a weighting scheme for volunteer panel web surveys '. *Journal of official statistics*, 22(2):329–349, 2006.

Lee, B.K., Lessler, J. and Stuart E.A. (2010) 'Improving propensity score weighting using machine learning '. *Statistics in medicine*, 29(3):337–346.

Lee, S.; Valliant, R. (2009). 'Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment '. *Sociological Methods & Research*, 37(3), 319-343.

Lohr, S. (2009) 'Multiple frame surveys '. In: Rao, C.R., Pfeffermann, D. (eds.) *Handbook of 24 Statistics, Vol. 29A, Sample Surveys: Design, Methods and Applications*, pp. 71–88. North Holland, Amsterdam

Lohr, S. and Rao, J. N. K. (2006). 'Estimation in multiple frame surveys '. *J. Amer. Statist. Assoc.*, 101(475), 1019 – 1030

Luebker, M. (2021). How much is a box? The hidden cost of adding an open-ended probe to an online survey. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *15*(1), 7-42.

Lyberg, L., Pennell, B.-E., Cibelli Hibben, K., de Jong, J. (2021). *AAPOR/WAPOR Task Force Report on Quality in Comparative Survey*s.

Lynn, Peter (2002) 'PEDAKSI: Methodology for Collecting Data about Survey Non-respondents ', *Working Papers of the Institute for Social and Economic Research paper 2002-05*. Colchester: University of Essex.

Madans, J., Miller, K., Maitland, A., & Willis, G. B. (2011). *Question evaluation methods: contributing to the science of data quality*. John Wiley & Sons.

Makstat (2023). *Active population in the Republic of North Macedonia - processed data from the Labour Force Survey*, *for 2021 and by quarters*.

Marken, S. (2018). *Still listening: The state of telephone surveys*. Gallup News. Mecatti, F. (2007). 'A single frame multiplicity estimator for multiple frame surveys ', *Survey Methodology*, 33 (2007), pp. 151–157.

Mecatti, F., Singh, A. (2014). Estimation in Multiple Frame Surveys: A Simplified and Unified Review using the Multiplicity Approach. Journal de la société française de statistique, 155(4), 51-69.

Meitinger, K. (2017). Necessary but insufficient: Why measurement invariance tests need online probing as a complementary tool. Public Opinion Quarterly, 81,447-472. doi: 10.1093/poq/nfx009

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

Metcalf, P., Scott, A. (2009). Using multiple frames in health surveys. Statistics in Medicine, 28(10), 1512-1523.

Miller, R. L., Collette, T. (2019). *Multicultural identity development: Theory and research. Cross-cultural psychology: contemporary themes and perspectives*, 614-631.

Molina, D., Rueda, M., Arcos, A., Ranalli, M.G., (2015) *Multinomial logistic estimation in dual frame surveys*, SORT 39, 2, pp. 309–336

Monstat (2022). *Labour Force Survey*. Release 37/2022.

OECD (2017), *OECD Guidelines on Measuring the Quality of the Working Environment*, Opsomer, J. D.; Erciulescu, A. L. (2021) 'Replication Variance Estimation After Sample-Based Calibration '. *Survey Methodology*, Statistics Canada Vol. 47 (No. 2).

Pasadas-del-Amo, S. (2018). 'Cell phone-only population and election forecasting in Spain: The 2012 regional election in Andalusia '. *Revista Espanola de Investigaciones Sociológicas (REIS)*, 162(162), 55-71.

Pennell, B-E., Cibelli Hibben, K. L., Lyberg, L., Mohler, P. Ph., 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, et al. (Eds.), *Total survey error in practice*. New York, NY: John Wiley & Sons.

Presser, S. et al. (2004). Public Opinion Quarterly. Methods for testing and evaluating survey questions, 68. Retrieved from https://doi.org/10.1093/poq/nfh008

Peytchev, A., Peytcheva, E. (2017). 'Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. In *Survey Research Methods (Vol. 11, No. 4, pp. 361-368)*.

Ranalli, M. G., Arcos, A., Rueda, M. and Teodoro, A. (2016). 'Calibration estimation in dual-frame surveys '. *Stat. Methods Appl.*, 25(3), 321-349

Riddles, M. K., Kim, J. K., Im, J. (2016). 'A propensity-score-adjustment method for nonignorable nonresponse '. *Journal of Survey Statistics and Methodology*, 4(2), 215-245.

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

Särndal, C.E.; Lundström, L. (2005). *Estimation in surveys with nonresponse*.

Scanlon, P. J. (2019). The effects of embedding closed-ended cognitive probes in a web survey on survey response. *Field methods*, *31*(4), 328-343.

Scanlon, P. (2020). Using targeted embedded probes to quantify cognitive interviewing findings. Advances in questionnaire design, development, evaluation and testing, 427-449.

Schneider,B. (2023) *svrep: Tools for Creating, Updating, and Analyzing Survey Replicate Weights*.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American sociological review*, 218-222

Seligson, M. A., Moreno Morales, D. E. (2018). 'Improving the Quality of Survey Data Using CAPI Systems in Developing Countries '. In L. R. Atkeson R. M. Alvarez (Eds.), *The Oxford Handbook of Polling and Survey Methods* (pp. 207–219). Oxford University Press.

Silva, P.L.D.; Skinner, C.J. (1997) *Variable selection for regression estimation in finite populations*. *Survey Methodology* 1997; 23: 23–32.

Singh, S., Sedory, S. A. (2016). 'Two-step calibration of design weights in survey sampling '. *Communications in Statistics-Theory and Methods*, 45(12), 3510-3523.

Smith, T. W. (2011). 'Refining the total survey error perspective '. *International Journal of Public Opinion Research*, 464–484.

Stoop, I; Billiet, J.; Koch, A.; Fitzgerald, R. (2010) *Improving Survey Response: Lessons Learned from the European Social Survey*

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.*

Sverchkov, M. (2008). 'A new approach to estimation of response probabilities when missing data are not missing at random '. In *Proceedings of the Survey Research Methods Section* (pp. 867-874).

Team, E. E. C. S. (2022). *Quality Report for the European Social Survey*, Round 9.

Technopolis Group (2022). *Sustain – 2: Impact study of the European Social Survey*

Thomas, S., Scheller, D., Schröder, S. (2021). 'Co-creation in citizen social science: The research forum as a methodological foundation for communication and participation '. *Humanities and social sciences communications*, 8(1), 1-11.

Tsung, C., Kuang, J., Valliant, R. L., Elliott, M. R. (2018). 'Model-assisted calibration of non-probability sample survey data using adaptive LASSO '. *Survey Methodology*, 44(1), 117-145.

Van de Vijver, F.; Leung, K. (1997). *Methods and data analysis of comparative research*.

Van Praag, L. (2021). 'Co-creation in migration studies: The use of co-creative methods to study migrant integration across European Societies '. *Leuven University Press*.

Verma V., (2014) *Sampling: An Introduction*

Verma, V., Betti, G. (2011). *Taylor linearization sampling errors and design effects for poverty measures and other complex statistics*. Journal of Applied Statistics, 38(8), 1549-1576.

Weisberg, H. F. (2009). 'The total survey error approach: A guide to the new science of survey research '. *University of Chicago Press*.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., Erikson, P. (2005). 'Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation '. *Value in health*, 8(2), 94-104.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. sage publications.

World Bank (2021). *Mobile cellular subscriptions (per 100 people) - European Union*.

World Health Organization (2022). Harmonized health facility assessment (HHFA): quick guide. Geneva.

Zavala-Rojas, D. (2014). *A procedure to prevent differences in translated survey items using SQP*.

7. Appendices

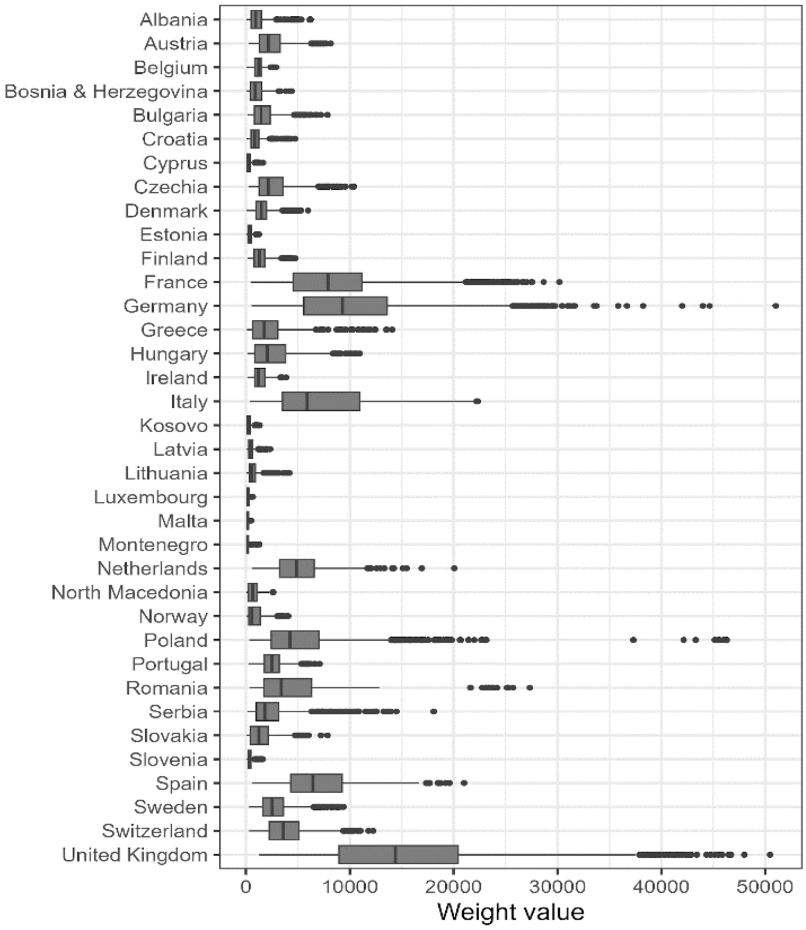
Table 23. Descriptive statistics of the final weights in every country

Country	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Std. deviation
Albania	50.45	494.79	925.23	1261.38	1456.67	6225.85	1253.68
Austria	260.85	1292.29	2154.40	2420.46	3278.12	8173.88	1479.16
Belgium	60.35	849.14	1216.13	1146.63	1447.84	2917.55	418.13
Bosnia & Herzegovina	27.23	435.30	890.67	1009.39	1491.45	4415.72	735.57
Bulgaria	144.01	807.41	1450.95	1712.97	2338.64	7833.42	1230.82
Croatia	75.46	485.11	830.96	932.33	1219.41	4731.49	630.42
Cyprus	24.68	123.02	252.18	316.26	412.50	1645.33	257.03
Czechia	232.35	1257.13	2151.97	2619.80	3549.58	10404.55	1840.69
Denmark	85.34	973.00	1488.62	1593.46	1983.97	5985.61	900.36
Estonia	28.52	205.72	334.77	362.64	510.52	1222.89	208.67
Finland	159.54	782.92	1268.64	1352.29	1820.19	4758.66	751.28
France	450.13	4570.28	7884.73	8629.82	11176.51	30180.59	5509.07
Germany	534.10	5575.68	9294.87	10045.99	13600.02	51010.61	6039.21
Greece	92.81	640.12	1769.68	2184.65	3062.94	14076.97	2015.99
Hungary	140.62	845.65	2054.29	2590.35	3836.14	10964.98	2133.25
Ireland	142.33	843.50	1179.43	1338.49	1830.29	3879.83	677.76
Italy	388.30	3504.51	5882.32	7203.45	10959.79	22327.21	4897.93
Kosovo	13.24	136.91	260.28	309.44	416.62	1322.44	238.10
Latvia	54.41	271.54	416.22	483.32	638.03	2325.02	290.77
Lithuania	83.49	354.20	551.35	731.48	890.72	4174.38	610.03
Luxembourg	18.23	144.59	213.32	225.09	287.50	652.48	107.21
Malta	19.05	119.35	176.94	182.20	235.33	506.19	82.75
Montenegro	9.20	94.80	160.48	185.19	241.02	1271.55	142.58
Netherlands	563.91	3224.25	4856.24	5111.34	6585.06	20069.59	2538.01
North Macedonia	29.85	232.31	665.22	699.30	1070.01	2633.36	519.25
Norway	43.60	273.63	559.40	847.27	1356.84	4071.39	701.13
Poland	330.50	2429.12	4243.50	5743.48	7037.63	46269.90	5509.16
Portugal	273.78	1784.09	2499.88	2559.73	3199.05	7089.91	1083.45

Country	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Std. deviation
Romania	347.47	1747.35	3392.76	4289.55	6332.62	27343.99	3570.91
Serbia	112.38	1008.97	1829.31	2479.37	3125.67	18112.04	2449.66
Slovakia	41.67	418.48	1233.50	1427.31	2149.05	7839.85	1213.21
Slovenia	37.65	232.39	349.99	369.29	474.53	1623.34	201.38
Spain	549.31	4313.02	6456.93	6811.44	9267.20	21002.24	3330.12
Sweden	301.46	1641.69	2524.35	2804.11	3592.08	9393.13	1622.25
Switzerland	310.36	2246.37	3588.85	3827.12	5094.80	12232.43	2100.92
United Kingdom	1255.15	8966.98	14382.92	15320.90	20437.01	50463.32	8385.27

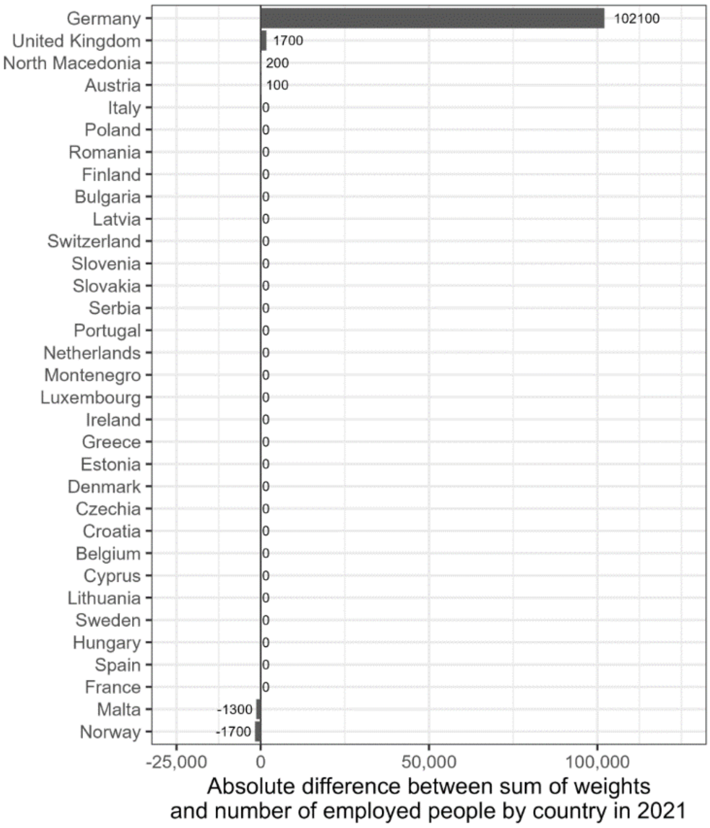
Source: Author's own elaboration

Figure 3. Boxplots of the final weights in every country



Source: Author's own elaboration

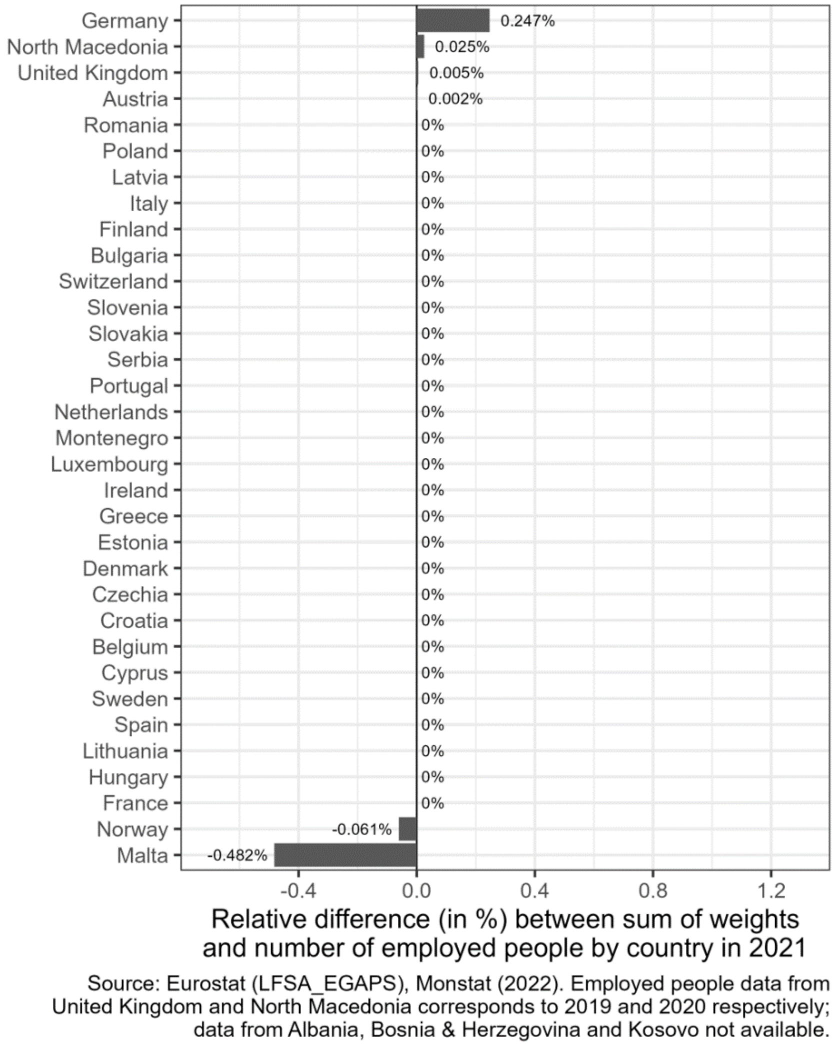
Figure 4. Absolute difference between sum of weights and number of employed people by country in 2021



Source: Eurostat (LFSA_EGAPS), Monstat (2022). Employed people data from United Kingdom and North Macedonia corresponds to 2019 and 2020 respectively; data from Albania, Bosnia & Herzegovina and Kosovo not available.

Source: Author's own elaboration

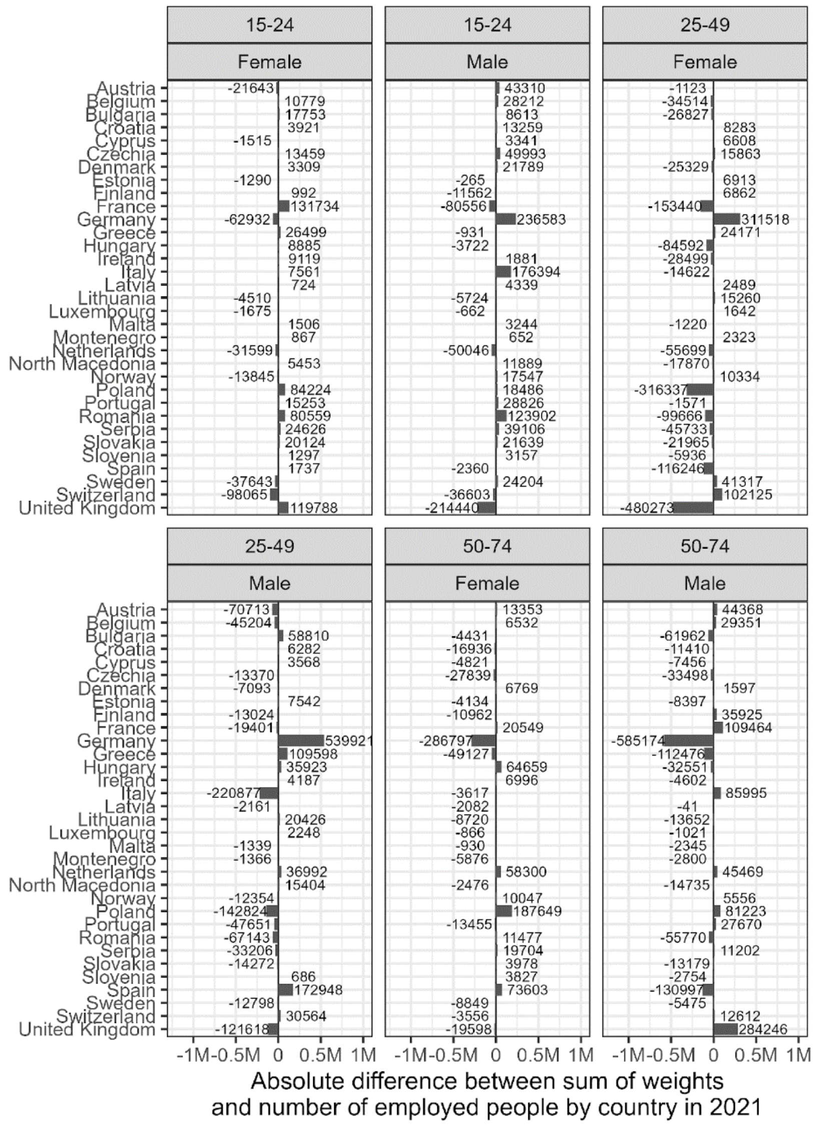
Figure 5. Relative difference between sum of weights and number of employed people by country in 2021



Source: Eurostat (LFSA_EGAPS), Monstat (2022). Employed people data from United Kingdom and North Macedonia corresponds to 2019 and 2020 respectively; data from Albania, Bosnia & Herzegovina and Kosovo not available.

Source: Author's own elaboration

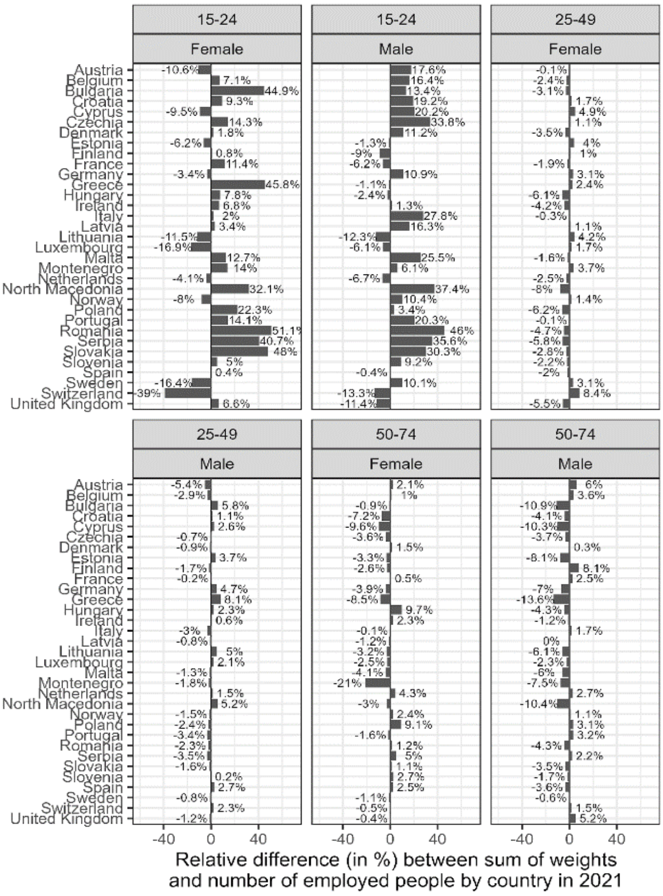
Figure 6. Absolute difference between sum of weights and number of employed people by country in 2021 by age and gender combined



Source: Eurostat (LFSA_EGAPS). Employed people data from United Kingdom, Montenegro and North Macedonia corresponds to 2019, 2020 and 2020 respectively; data from Albania, Bosnia & Herzegovina and Kosovo not available.

Source: Author's own elaboration

Figure 7. Relative difference between sum of weights and number of employed people by country in 2021 by age and gender combined



Source: Eurostat (LFSA_EGAPS). Employed people data from United Kingdom, Montenegro and North Macedonia corresponds to 2019, 2020 and 2020 respectively; data from Albania, Bosnia & Herzegovina and Kosovo not available.

Source: Author's own elaboration

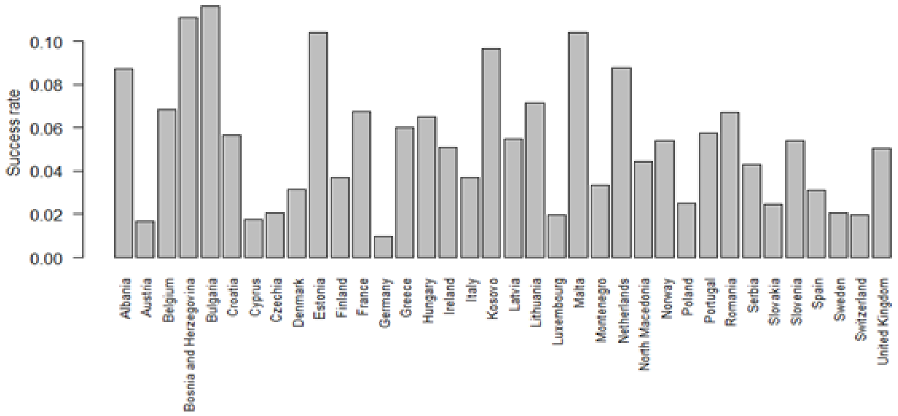
Table 24. Regression coefficients, standard errors, t values and p-values for the regression model to predict the design effect according to age, gender, sector and occupation

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.063	0.505	6.064	0.000001
Median age	-0.041	0.014	-2.888	0.007
Proportion of female	-0.205	0.71	-0.289	0.774
Proportion of highly skilled workers	0.191	0.676	0.283	0.78
Proportion of "white-collar" workers	0.379	1.044	0.363	0.719

Adjusted R^2 = 0.1725

Source: Author's own elaboration

Figure 8. Yield by countries[2]



Source: Author's own elaboration

Table 25. Percentage of outliers

Variable	Question	Outliers	Percentage of outliers
time_care_children_minutes	Q96C. On average, how many hours per day do you spend on the activity? C. Caring for and/or educating your children, grandchildren (please, also include people, who do not live in the household)	536	0,7367
time_housework_minutes	Q96D. On average, how many hours per day do you spend on the activity? D. Cooking and housework	2452	3,3701
time_care_relatives_minutes	Q96E. On average, how many hours per day do you spend on the activity? E. Caring for elderly/ disabled relatives (please, also include people, who do not live in the household)	255	0,3505

Source: Author's own elaboration

Table 26. Percentage outliers by country

	time_care_children_minutes	time_housework_minutes	time_care_relatives_minutes
Albania	1,3145	2,7300	0,9100
Austria	0,6745	2,5857	0,1124
Belgium	1,2521	4,4649	0,4252
Bosnia & Herzegovina	0,8772	1,0526	0,4386
Bulgaria	0,6682	2,6169	0,4454
Croatia	0,7778	1,6667	0,7778
Cyprus	0,4396	1,0989	0,0733
Czechia	0,5528	2,0101	0,1508

[2] Final number of interviews achieved after all quality checks/actual gross sample.

Denmark	0,4396	5,9890	0,0549
Estonia	0,2217	3,8248	0,5543
Finland	1,5239	8,3027	0,4204
France	0,6847	5,7890	0,3735
Germany	0,2179	2,4207	0,1452
Greece	0,9455	2,7253	0,3893
Hungary	0,7254	2,9018	0,3906
Ireland	1,0644	4,3697	0,3361
Italy	0,7985	1,8524	0,2874
Kosovo	0,0882	0,2646	0,0000
Latvia	1,0006	3,3352	0,3891
Lithuania	0,3741	1,8172	0,3207
Luxembourg	0,4402	4,1086	0,1467
Malta	0,3397	1,8342	0,4076
Montenegro	0,2613	0,4355	0,6969
Netherlands	1,0463	5,2863	0,4405
North Macedonia	4,0457	5,0132	2,4626
Norway	0,9697	8,3939	0,303
Poland	0,4483	1,8966	0,0345
Portugal	0,4787	3,1915	0,2128
Romania	0,5531	1,2168	0,5531
Serbia	0,6092	2,9591	0,8703
Slovakia	1,1148	2,1182	0,223
Slovenia	0,9122	2,8506	0,4561
Spain	0,3789	2,5835	0,2067
Sweden	0,7119	5,6407	0,1643
Switzerland	0,4902	2,8595	0,0817
United Kingdom	0,4217	3,5145	0,1406

Source: Author’s own elaboration

Table 27. Error of the interest variables

Interest variable	Standard error	Variation coefficient
employee_selfdeclared (Employee)	0,0049	0,0057
private_sector (The private sector)	0,0046	0,0071
usual_days	0,0372	0,0076
commute_days	0,0393	0,0086
usual_hours_week	0,2236	0,0057
seniority	0,0943	0,0083

contract_duration_month	0,4455	0,0312
chemicals (Sometimes)	0,0022	0,0219
infect (Sometimes)	0,0018	0,0271
night (Sometimes)	0,0019	0,0198
asb_verbal (Yes)	0,0025	0,0257
asb_unwanted_sexatt (Yes)	0,0012	0,0600
asb_violence_harassment (Yes)	0,0019	0,0316
osh_risk (Yes)	0,0035	0,0108

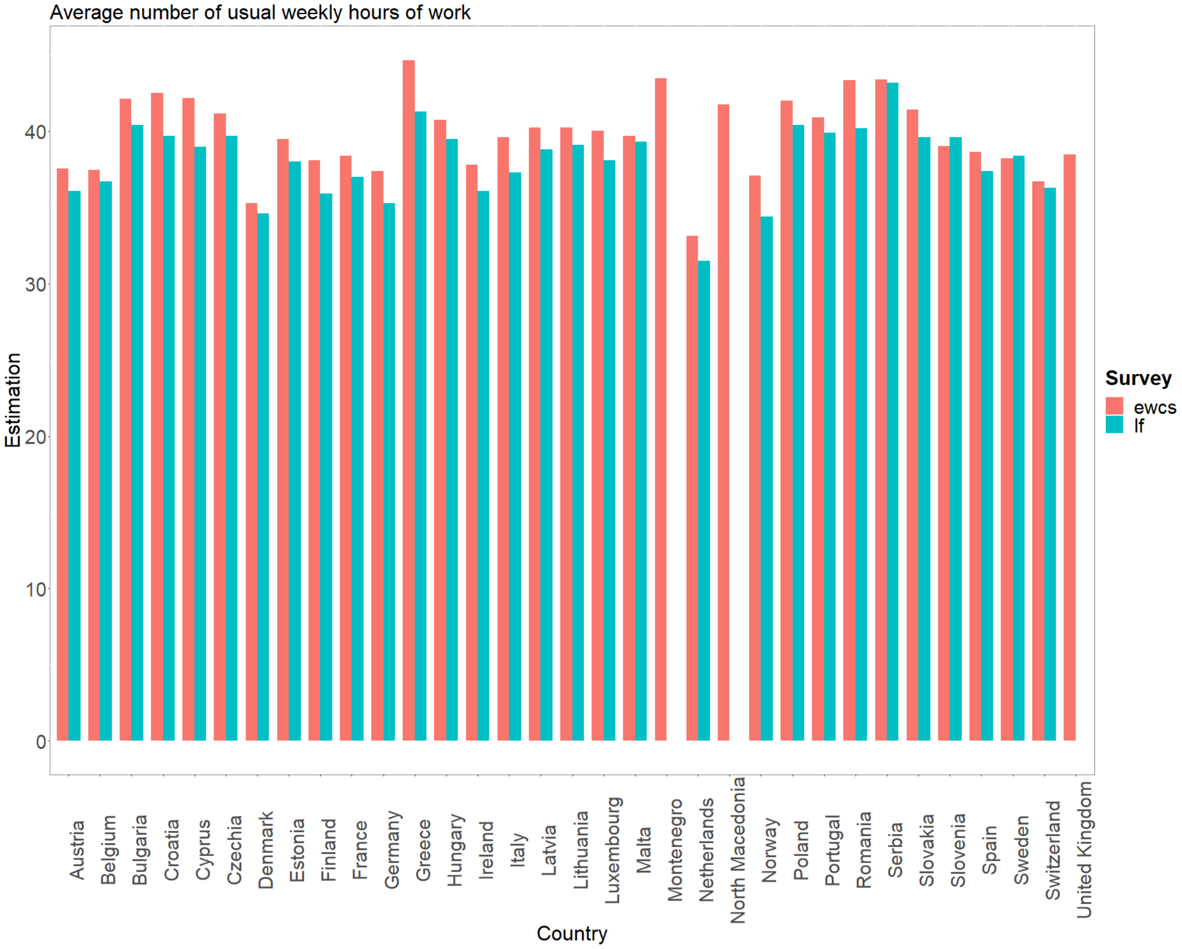
Source: Author’s own elaboration

Table 28. Design effects by country

Country	Design effects
Austria	1,3735
Belgium	1,1330
Bulgaria	1,5163
Croatia	1,4572
Cyprus	1,6605
Czechia	1,4937
Denmark	1,3193
Estonia	1,3311
Finland	1,3087
France	1,4075
Germany	1,3614
Greece	1,8516
Hungary	1,6782
Ireland	1,2564
Italy	1,4623
Latvia	1,3619
Lithuania	1,6955
Luxembourg	1,2268
Malta	1,2063
Montenegro	1,5928
Netherlands	1,2466
North Macedonia	1,5514
Norway	1,6848
Poland	1,9201
Portugal	1,1792
Romania	1,6930
Serbia	1,9762
Slovakia	1,7225
Slovenia	1,2974
Spain	1,2390
Sweden	1,3347
Switzerland	1,3014
United Kingdom	1,2995

Source: Author’s own elaboration

Figure 9**. LFS estimation versus EWCS 2021 estimation by country**



Source: Autor's own elaboration

WPEF24067

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency established in 1975. Its role is to provide knowledge in the area of social, employment and work-related policies according to Regulation (EU) 2019/127.

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