

Social tracking methodology for the EU budget



Social tracking methodology for the EU budget

Abstract

The EU tracks social expenditure for several reasons, ranging from ensuring financial control to assessing whether policies make the desired impact. This study:

- examines the **current state of social tracking** in the Multiannual Financial Framework and the Recovery and Resilience Facility;
- explores **how it could be improved** to enhance the scrutiny of EU social spending;
- puts forward a **proposal for a pilot study** of a theory-based monitoring approach as a potential innovation in social tracking.

This document was requested by the European Parliament's Committee on Budgets.

AUTHORS

Iain BEGG, LSE and CEPS
Francesco CORTI, CEPS
Alessandro LISCAI, CEPS
Tamás KISS-GÁLFALVI, CEPS
Tommaso GROSSI, EPC
Laura RAYNER, EPC

ADMINISTRATOR RESPONSIBLE

Stefan SCHULZ

EDITORIAL TEAM

Eleanor JAMES, Adrienn BORKA

LINGUISTIC VERSIONS

Original: EN

ABOUT THE EDITOR

Policy departments provide in-house and external expertise to support EP committees and other parliamentary bodies in shaping legislation and exercising democratic scrutiny over EU internal policies.

To contact the Policy Department or to subscribe for updates, please write to:

Policy Department for Budgetary Affairs

European Parliament

B-1047 Brussels

Email: Poldep-Budg@ep.europa.eu

Manuscript completed in January 2023

© European Union, 2023

This document is available on the internet at:

<http://www.europarl.europa.eu/supporting-analyses>

DISCLAIMER AND COPYRIGHT

The opinions expressed in this document are the sole responsibility of the authors and do not necessarily represent the official position of the European Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

© Cover image used under licence from Adobe Stock

CONTENTS

LIST OF ABBREVIATIONS	5
LIST OF BOXES	6
LIST OF FIGURES	6
LIST OF TABLES	6
EXECUTIVE SUMMARY	7
INTRODUCTION	11
1. PERFORMANCE-BASED BUDGETING AND SOCIAL MAINSTREAMING	12
1.1. Performance-based budgeting	12
1.2. Budgetary mainstreaming	15
2. OVERVIEW AND BENCHMARKING OF THE EXISTING EU SOCIAL TRACKING SYSTEMS	17
2.1. Monitoring and Evaluation (M&E) systems in place for the ESF, YEI and FEAD	18
2.1.1. Monitoring	18
2.1.2. Evaluation	19
2.2. Shortcomings and limitations of the ESF M&E systems	20
2.3. Monitoring and Evaluation (M&E) in the ESF+: key novelties	22
2.4. Monitoring and Evaluation (M&E) in the RRF	24
2.4.1. Limits of M&E of the RRF	25
3. EXPLORING ALTERNATIVE APPROACHES	27
3.1. Theory-based evaluation (TBE) and its application in the current M&E framework	27
3.2. The advantages and drawbacks of TBE	29
3.3. A blind spot: the 'black box' approach to causal pathways	30
3.4. Potential methods to measure impacts and policy utility: contribution analysis and contribution tracing	31
3.4.1. Overview of contribution analysis and contribution tracing	31
3.5. Practical applicability	33
3.6. Towards implementation of a TBE approach to social tracking	35
4. MAINSTREAMING SOCIAL OBJECTIVES IN EU SPENDING	37
4.1. Existing EU mainstreaming objectives practiced in the EU budget	37
4.1.1. Climate mainstreaming	37
4.1.2. Gender mainstreaming	39

4.2.	Mainstreaming the EPSR objectives in the EU budget	41
4.2.1.	Debate on the EU Social Taxonomy	41
4.2.2.	Application of the AAAQ criterion to EU social spending	44
4.2.3.	Application of the social DNSH principle to EU spending	44
5.	CONCLUSIONS AND RECOMMENDATIONS	47
5.1.	Existing monitoring and evaluation framework	48
5.2.	Data and indicator problems	49
5.3.	Mainstreaming social objectives in the EU budget	50
5.4.	Recommendations	51
5.4.1.	Strengthen the current M&E system to improve social tracking	52
5.4.2.	Launch a pilot project on theory-based evaluation of selected social policies	53
	REFERENCES	55
	LIST OF INTERVIEWS	60

LIST OF ABBREVIATIONS

AAAQ	Availability, Accessibility, Acceptability, and Quality
AIRS	Annual Indicator Report Series
AROEPE	At Risk of Poverty or Social Exclusion
CA	Contribution Analysis
CC-ME	Competence Centre on Microeconomic Evaluation
CIA_s	Counterfactual Impact Assessments
CID	Council Implementing Decision
CIE	Counterfactual Impact Evaluation
CPR	Common Provisions Regulation
CRIE	Centre for Research on Impact Evaluation
CT	Contribution Tracing
DNSH	Do No Significant Harm
EC	European Commission
ECA	European Court of Auditors
ECEC	Early Childhood Education and Care
ESF	European Social Fund
ESF+	European Social Fund Plus
ESI	European Structural and Investment (Funds)
EP	European Parliament
EPSR	European Pillar of Social Rights
ERDF	European Regional Development Fund
EU	European Union
FEAD	Fund for European Aid to the most Deprived
IIA	Interinstitutional Agreement
ILO	International Labour Organisation
IL	Intervention Logic
JRC	Joint Research Centre
MA_s	Managing Authorities
M&E	Monitoring and Evaluation
MFF	Multiannual Financial Framework
M&T_s	Milestones and Targets
NACE	Statistical classification of economic activities in the European Community (French: Nomenclature statistique des activités économiques dans la Communauté européenne)
NEET	Not in Education, Employment or Training
NRRPs	National Recovery and Resilience Plans
OECD	Organisation for Economic Co-operation and Development
OP	Operational Programme
OSH	Occupational Safety and Health
RRF	Recovery and Resilience Facility
SCO	Simplified Cost Option
SDGs	Sustainable Development Goals
TBE	Theory-Based Evaluation
TBM	Theory-Based Monitoring
YEI	Youth Employment Initiative

LIST OF BOXES

Box 1: ERDF social objectives and key common output and result indicators	17
Box 2: The Green Taxonomy	39

LIST OF FIGURES

Figure 1: System view of an initiative	14
Figure 2: Steps in Contribution Analysis	32
Figure 3: Steps in Contribution Tracing	33

LIST OF TABLES

Table 1: Applicability of contribution analysis and contribution tracing to EU M&E	35
Table 2: Application of Social Taxonomy to adequate living standards and wellbeing for users (objective), notably in access to healthcare (sub-objective)	42
Table 3: Application of Social Taxonomy to adequate living standards and wellbeing for users (objective), notably in access to housing (sub-objective)	43

EXECUTIVE SUMMARY

The objective of this study is to examine the existing means by which expenditure from the EU budget for social policies is monitored and tracked. Effective tracking is important for policy purposes to enable programme progress to be assessed, and to be able to detect, qualify and quantify any deviation from the initial plans and targets.

Pressures for better regulation led to a range of new approaches to appraising the quality of budgetary oversight, notably the shift to performance-based budgeting. The adoption of performance budgeting implies a shift in the focus of budgeting, from managing (i.e. 'how much have we spent?'), to the achievement of policy objectives (i.e. 'what have we accomplished with our money?').

The notion of the 'logic of intervention' is, therefore, central to providing answers, because it enables political decision-makers to conceptualise how they envisage a policy should work and then assess whether it fulfils these expectations.

Existing systems of tracking

The existing systems for monitoring spending from the EU budget have evolved considerably, especially in the course of the 2014-20 programming period and with the changes subsequently consolidated in the current 2021-27 Multiannual Financial Framework (MFF). The main thrust of these innovations was to require more of a performance orientation and to make ex post evaluations mandatory across the policy cycle.

Although existing systems have tried to change the emphasis in monitoring towards the results of programmes and the impact they are expected to have, for the most part the principal change has been to give more prominence to direct outputs of programmes. Consequently, too little information is, or will be, generated about whether EU spending succeeds in stimulating societal transformations.

Furthermore, data collection is still perceived as an administrative burden by implementing authorities and not as an exercise to improve the quality of the policies. As a result, the usefulness of some indicators is questioned, especially those subject to lags in collecting them, because the information generated may already be outdated.

By comparison with other EU spending programmes for which hard data can readily and quickly be collected, specific difficulties arise in tracking the impacts of social spending for three main reasons:

- First, social change inherently takes time: unambiguous impacts frequently only become visible well beyond the life of a programme. In addition, many of the indicators used rely on surveys or other data collection techniques which only yield findings after a longer lag.
- Second, spending from the EU budget is only one of many funding streams affecting social targets, making it hard to distinguish its contribution, especially when the amounts disbursed are small compared to those at national and (sometimes neglected) sub-national levels. In some circumstances, however, EU expenditure may have filled a gap not covered by support from national or other sources.
- A third problem is determining causality where social change may well have occurred regardless of policy interventions. Put succinctly: correlation is not necessarily causality.

In addition, the most recent efforts to improve EU performance budgeting have focused on improving evaluations, while the attention devoted to monitoring has been only functional to the preparation of the evaluation phase. This limits the potential of the monitoring exercise, which, in performance-based

systems, should also include detecting deviations from the expected implementation plan and pointing towards how to correct them.

The Recovery and Resilience Facility (RRF) introduced a new performance-based approach which links funding disbursement to the achievement of milestones and targets (M&Ts). A ‘financing not linked to costs’ approach is not a novelty in EU budgeting and was first formally introduced as a possible alternative to the Simplified Cost Option (SCO) in the Financial Regulation of the EU budget in 2018. Yet, with the RRF it becomes the rule. Even though the RRF is a step forward in the current Monitoring and Evaluation (M&E) framework, it still has some limitations:

- First, the selected M&Ts largely rely on output indicators, and often these are presented in an aggregated way;
- Second, the targets largely monitor infrastructural outputs and ignore societal outcomes; and
- Third, the milestones are sometimes too vague on the exact commitment of national governments when it comes to the implementation of reforms.

In addition to these three limitations of the current M&Ts, a more structural problem of the RRF in its current form lies in its rigidity, which does not allow for early detection of deviation from the expected results and correction of the same. M&Ts are indeed set *ex ante*. Due to inherent uncertainty and rapidly changing conditions in policymaking, this is no easy task. Modern investment practice often relies on smaller milestones and regular joint review to allow for revision. This is however not true of the RRF.

A theory-based approach to social tracking

Given the limitations of the current approach to social tracking, this study has explored alternatives. It is, however, important to stress that expectations of what can be achieved have to be realistic and that there can be no ideal solution. However, a potentially promising option is to adapt Theory-Based Evaluation (TBE) for monitoring purposes. TBE is a demanding methodology and typically involves complex and costly research. Yet by emphasising the centrality of the ‘logic of intervention’ as a basis for monitoring progress, this study maintains that the use of TBE can produce valuable findings.

Building on the TBE approach, EU social spending would be assessed not just based on the annual/biannual achievement of M&Ts or the costs incurred, but rather on the qualitative assessment of the effectiveness and utility of the interventions. This requires a combination of data collection on monetary inputs at the disaggregated level to account for the cross- and intra-regional distribution of resources, and significant *ex ante* knowledge of what to expect from social expenditure.

Arriving at a coherent understanding of projected socio-economic returns calls for awareness of the micro-foundations of social change (i.e. micro-outcomes in life course transitions), complementarities and sequencing of reform trajectories, and understanding of the institutional mechanisms which have a bearing on national and subnational delivery of social policies.

Such an approach would yield several potential benefits. For national stakeholders, there would be some easing of the current administrative burden in reporting on output indicators, limiting problems in defining not only common indicators, but also outputs and results. In parallel, it would help to increase stakeholder involvement in programme monitoring and allow for flexible adjustments.

For the European Commission it would deter unproductive box-ticking in monitoring the implementation of programmes, but also pave the way for mid-term evaluations of the MFF/RRF, while also increasing transparency and flexibility.

Mainstreaming of social policy

As a complementary avenue to social tracking, budgetary mainstreaming is about how policies, programmes and actions are formulated, designed, and implemented and monitored, taking account of overarching societal goals and priorities. This study explores the possibility to introduce social mainstreaming practices to make sure that all EU-funded policies are compliant with the social standards defined in the European Pillar of Social Rights (EPSR).

Building on the current debate about the establishment of an EU Social Taxonomy, two different approaches to social mainstreaming are identified: one vertical and one horizontal. The first is about guaranteeing that EU-funded social spending is compliant with common EU social objectives. The second concerns making sure that non-social EU spending does not significantly harm EU social objectives.

In the first case, the study explores the possibility of applying the availability, accessibility, acceptability, and quality (AAAQ) criteria to all social spending. Applying these criteria should be done in two steps. Social spending under the EU budget should be compliant with the four criteria and their approval should be conditional on this. Data and information should be properly collected, drawing on and consulting relevant stakeholders. Monitoring and tracking arrangements should be defined ex ante. Regular monitoring is essential for vertical social mainstreaming to ensure that all the participants involved are aware of the relevant objectives and plans or, if this is not the case, that there is an obligation to set up briefings, capacity-building meetings and appropriate training.

In the second case, the study explores how a social Do Not Significantly Harm (DNSH) procedure would be applied horizontally to all EU spending. Concretely, this means that any EU spending should be checked to verify that it does not violate social objectives. The study identifies, based on the Social Taxonomy proposal, three objectives to be monitored (the approach could be further expanded), reflecting the principles of EPSR Chapter 2 – ‘Fair working conditions’: promoting decent work; promoting equality and non-discrimination at work; and ensuring respect for the human rights and workers’ rights of affected workers in the value chain.

The study illustrates the pros and cons of both alternatives and concludes that there is considerable potential for developing social mainstreaming, based on the principles of the EPSR, as a tool for all headings of expenditure. However, the way in which to do so remains contentious; the study thus recommends exploring in more detail how mainstreaming might be made both politically acceptable and possible to implement in a credible manner.

Key recommendations

The study puts forward recommendations for how the European Parliament (EP) can improve social tracking. These recommendations are pitched at two levels: strengthening the current M&E system to improve social tracking and launching a pilot project on theory-based monitoring of selected social policies. Clarity on the logic of intervention is critical, because the concern of the EP as a budgetary authority is not only budget execution, but also whether the policies adopted fulfil political objectives and demonstrably have EU added value.

Correlation is not causality and a cautionary conclusion is that tracking may not yield convincing answers, hence expectations for the quality of information that can be obtained should remain realistic.

The contribution of EU financing to social objectives cannot be fully disentangled from national level funding programmes.

To strengthen the current system, the study proposes to:

- Aim for greater coherence in the definitions of key concepts and indicators by ensuring, as far as possible, consistency in definitions across EU legislation related to expenditure programmes;
- Enable managing authorities to access administrative data for monitoring and evaluation purposes, thereby reducing the administrative burden on intermediaries and beneficiaries of EU funds, and, in turn, allowing for more detailed analyses to be produced and for an improved evidence base for future policymakers;
- Guarantee that regional and local actors are better engaged in the monitoring and evaluation process, and are involved in the identification of territory-specific indicators to track social spending and evaluate performance;
- Give priority to developing a framework for learning evaluation in order to foster greater consistency in data and improve the exchange of experience in social tracking.

Under an alternative approach, the study proposes the following:

- A pilot project for the theory-based monitoring of selected social policy areas in the RRF should be launched. This entails agreeing the 'logic of intervention'.
- The selected programmes should be assessed not based on common indicators or the annual/biannual achievement of M&Ts, but rather on the qualitative assessment of the effectiveness and utility of the interventions.
- An essential part of the exercise will be for the EP to set out its own logic of intervention for the policy in question, thereby enabling it to identify the causal mechanisms expected to lead to results and impacts.
- Progress can then be monitored on a separate track from the formal monitoring of M&Ts required for disbursements of tranches of funding, with the focus of the exercise being on whether the policy 'works'.
- A multi-stakeholder approach – with the involvement of sub-national actors in charge of implementation – should be adopted.
- The results of this pilot exercise should inform the mid-term revision of the RRF.

INTRODUCTION

Expenditure on social policy initiatives is a sizeable proportion of the EU budget and has acquired added political weight following the adoption of the EPSR as a core, values-based dimension of the EU as a polity. For the EP, as a budgetary authority, it is important to know whether the funding provided is not only being spent properly from the perspective of financial control, but also whether it is contributing sufficiently towards realising the objectives and targets of the EPSR. Social tracking is the means by which the EP can monitor progress.

Although a scoreboard comprising 14 headline and 21 secondary indicators (subsequently enhanced at the 2021 Porto Social Summit) was established to monitor employment and social developments in participating Member States, its reliance on quantitative data limits its usefulness for tracking purposes. It is no surprise in this respect that – despite the explicit link between the EU budget and the principles of the EPSR – no methodology has been identified to track EU spending social outcomes and impacts. The objective of this study is to examine the features of current tracking practices and methodologies, and to explore options for improving how tracking is undertaken.

Unlike many projects undertaken as part of EU regional policy, for which physical and other quantitative indicators can provide valuable insights – although even then, data on roads, bridges and so on do not necessarily reveal how transformative a policy was – advances towards many social goals are much less easily captured by simple indicators. Establishing a system able to shed light on underlying changes is not an easy task. Social objectives and targets are intrinsically long-term because they involve transformations intended to alter the nature of society. They are also determined by a wide range of policy interventions: some are explicitly targeted at identifiable social outcomes, while others have a more indirect effect.

For policy purposes, including monitoring progress while a programme is underway, as well as feeding into future policy, effective tracking has to be able to detect, qualify and quantify any deviation from the initial plans and targets. It also has to reconcile the desire for extensive information with breaching data protection rights and other sensitivities of beneficiaries and avoiding an excessive administrative burden on public bodies implementing policies.

To contribute to improvements in social tracking, this study is composed of several research tasks. Task 1 is to review the conceptualisation of two key elements, performance-based budgeting and mainstreaming, with the purpose of enriching the theoretical framework for social tracking and drawing out implications for methodologies. Task 2 assesses the monitoring and evaluation system in place for the European Social Fund Plus (ESF+) during the 2021-2027 period, and compares it with the corresponding arrangements for the previous period in which differing systems were applied to the European Social Fund (ESF) and certain other expenditure programmes.

Task 4 consists in an exploration of how social tracking might be enhanced, focusing mainly on the scope for building on the notion of the ‘logic of intervention’. The study looks at the feasibility of adapting theory-based evaluation (TBE) for monitoring purposes. TBE calls for the identification of the theories underpinning the rationale, causal mechanisms, interrelations and objectives of different forms of policy intervention. The approach is well suited to circumstances in which what would happen in the absence of the policy (the ‘counterfactual’) cannot easily be ascertained, something that characterises social intervention for which impact – in the sense of societal transformation – often takes a long time to be felt. Task 4 is to survey whether, and how, social objectives enshrined in the EPSR are mainstreamed in the design, preparation, implementation and evaluation of each EU spending programme. The last chapter presents the conclusions of the study and puts forward recommendations for how the EP can improve social tracking.

1. PERFORMANCE-BASED BUDGETING AND SOCIAL MAINSTREAMING

Performance budgeting is a concept that, in recent years, has gained traction at EU level. Already in 2015, the European Commission launched the initiative *'The EU budget focused on results'*, centred on fund performance, efficiency and effectiveness. Starting with the MFF 2014-2020, a 'logic of intervention' approach was introduced in the EU budget, notably in EU cohesion policy, with the aim of shifting monitoring and evaluation from disbursement procedures and costs to results. Despite this new focus on performance, no mandatory procedure was adopted. The upshot was that the performance-based approach was left as optional, with the evidence indicating that it had been adopted by hardly any of the Member States.

Similarly, the mandatory performance reserve was released on the basis almost exclusively of progress in spending and outputs, not outcomes. The launch of the RRF, as will be illustrated in more detail in the next section, led to renewed debate on performance budgeting and on how to track the performance of EU spending. The purpose of this section is thus to analyse, from a theoretical perspective, what is meant by performance-based budgeting and what proper tracking of EU spending might entail.

This section then examines another increasingly relevant concept related to EU spending, i.e. budgetary mainstreaming. While mainstreaming and tracking are complementary concepts, they should be analytically distinguished. This has also a consequence methodologically, since the implementation of policy mainstreaming in budgetary policies automatically implies double counting of expenditure (European Commission, [2022a](#)). This is not an error but a feature of mainstreaming. In what follows, the existing academic and grey literature on budgetary mainstreaming is investigated.

1.1. Performance-based budgeting

The idea of performance-based budgeting was conceived in the early 2000s, within a debate on how to measure the results (outputs and outcomes) of public spending (Sapała, [2018](#)). The Organisation for Economic Co-operation and Development (OECD) has provided the most comprehensive definition of this budgetary method, describing it as *'the systematic use of performance information to inform budget decisions, either as a direct input to budget allocation decisions or as contextual information to inform budget planning, and to instil greater transparency and accountability throughout the budget process, by providing information to legislators and the public on the purposes of spending and the results achieved'* (Schick, [2014](#)). Unlike the traditional budgeting approach, performance budgeting aims at allocating resources where the objectives are more likely to be achieved. In other words, the adoption of performance budgeting implies a shift in the focus of budgeting from managing (i.e. 'how much have we spent?'), to the achievement of policy objectives (i.e. 'what have we accomplished with our money?') (OECD, [2008](#)).

A crucial prerequisite for the success of performance budgeting is the collection of systematic, simple, affordable and usable performance information (Robinson and Last, [2009](#)). To this end, monitoring and evaluation systems, defining indicators for public spending outputs and outcomes, could be useful. They allow for assessments that turn performance information into policy recommendations which can then be implemented, allowing governments to monitor the performance of non-financial aggregates (Krause, [2010](#)).

The Better Regulation Guidelines – which summarise the principles followed by the European Commission for preparing, managing and evaluating legislation and policy initiatives – state that 'monitoring is a continuous and organised process of systematic data collection (or access) throughout

the life-cycle of any initiative to oversee its progress' (Tool #43). Monitoring as such looks at the 'what': what has occurred; what has been the output of an intervention. Building on this, a policy evaluation judges the extent to which the intervention is effective, efficient, relevant, coherent and brings added value compared to prospective alternatives. Evaluations go beyond a factual assessment of the 'what' to consider why it has happened, how much change can be attributed to the EU intervention and to what extent this change meets the original expectations or projections. Evaluations aim in principle to draw conclusions about the causal effects of the EU intervention on the actual outcomes/results (Tool #45). Monitoring and evaluation are thus complementary, as they ultimately aim to check whether policy objectives are being achieved.

Accordingly, the prerequisite for any performance-based budgeting is an appropriate monitoring arrangement that needs to be set up for collecting, processing and using/reusing data. Indicators are the main instruments for monitoring, since they track measurable progress towards defined objectives¹. As such they should be clearly defined and closely linked to the activity. The measurement unit selected should be defined precisely, and indicators have to be measured sufficiently frequently to monitor progress. Four types of indicators to track spending can be distinguished:

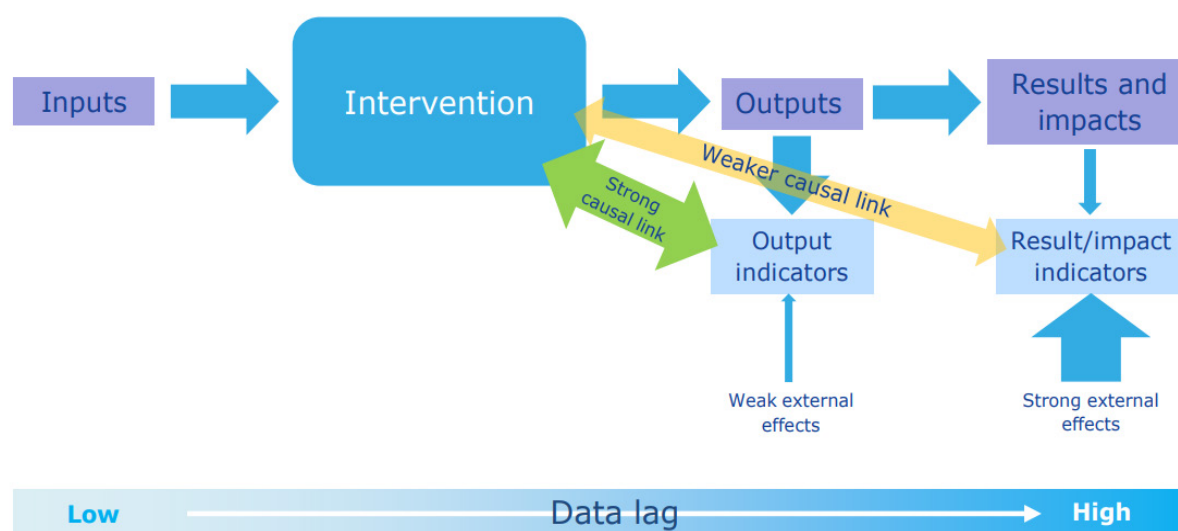
- **Input indicators** traditionally relate to the expenditure planned and disbursed, and thus the take-up of resources;
- **Output indicators** measure the immediate deliverables of the intervention, produced as a consequence of the initiative. They relate to what is directly produced and supplied through the implementation of financially supported operations;
- **Result indicators (also known as outcome indicators²)** measure the effects of the interventions supported, with particular reference to the direct addressees, population targeted or users;
- **Impact indicators** refer to the long-term impact on participants or broader target groups in society, beyond the participants or entities directly benefiting from support.

As observed in the Better Regulation Toolbox (European Commission, [2021a](#)), the more we move towards impact indicators, the more difficult it is to link the initiatives to what they deliver as results and impacts. This is due to a variety of factors, making it difficult to differentiate between correlation, causality and incidental effects. Moreover, data lags tend to be longer for impact indicators. For this reason, a proper monitoring system should also collect contextual information to account for developments that are not intentionally related to the individual initiative, but can influence it. From the perspective of policymakers, indicators capturing results and impacts are the most important to monitor, yet it can be difficult to find a causal link between the intervention and its impact. In addition, depending on the initiative, it can take some time before it has an impact and even more to be able to measure the impact with an indicator. Causality and timeliness are thus two aspects that complicate any monitoring and evaluation exercise, as illustrated in the figure below.

¹ The Better Regulation Toolbox (European Commission, [2021a](#)) provides clear criteria for identifying the indicators, which should be relevant, acceptable, credible, easy and robust (RACER). Additional criteria are, for instance, that indicators should be attributable to the initiative, data should be easily available and of good quality, indicators should capture the effects due to the initiative within reasonable timescales, and the baseline and targets should be defined ex ante.

² In EU terminology, the word 'results' is used in a precise sense to capture effects beyond direct outputs. The term 'outcome', more common in the wider evaluation literature, is used interchangeably with 'results' in this study. A difficulty with using 'outcome' is that it does not translate readily into other languages, except as 'result': thus, in Italian, it would be 'risultato'.

Figure 1: System view of an initiative



Source: Better Regulation Toolbox (European Commission, 2021a), p. 361.

When it comes to social measures, tracking the results and impacts of social spending is particularly difficult. Iliana Ivanova, European Court of Auditors (ECA) Member and Chair of ECA Chamber II – Investment for cohesion, growth and inclusion, expresses it well: ‘in terms of performance [the social policy area] is more difficult to measure because the results are more intangible’ (ECA, 2022). Several of the experts interviewed for this study have made similar points. Samson et al. (2015) argue that ‘social policy impact evaluation falls within the realm of activities in which “acts of choice” have influenced pathways to an outcome’.

As stated by Amartya Sen (2009), such acts fall within the scope of comprehensive outcomes which derive from processes of agents, institutions and relations among actors. This should be distinguished from culmination outcomes, relying on a clear itemisation of inputs that, when applied to a given issue, generate certain measurable impacts in a linear input-output causality (Arrow, 2006). In particular, the comprehensive evaluation of a social policy should assess the impact of such a policy on several outcomes and on the synergies between them (Samson et al., 2015), a task which is difficult to accomplish. Moreover, the evaluation has to be rigorously replicated in a variety of contexts in order to verify the consistent impact pathways of a specific intervention (Acemoglu, 2010; Devereux and Roelen, 2014: 1-2). This forces stakeholders to acknowledge that the effect of a policy may not be generalised (Samson et al., 2015).

An additional problem that emerges when tracking social policy outcomes concerns the nature of social indicators. They can be subject to high data volatility, time lags and a lack of data disaggregation (Istituto per la Ricerca Sociale, 2016). If one takes the EPSR scoreboard, all the indicators included are outcome indicators. For instance, the At Risk Of Poverty or Social Exclusion (AROPE) rate only becomes available after a significant delay and partly relies on subjective information. The definition of poverty remains difficult, as it does not measure ‘absolute’ poverty, but rather income inequality, because it is calculated relative to a median population value.

Similarly, severe material deprivation, one of the components of AROPE, presents problems in terms of comparability since the basic items included in the definition vary across countries. Youth unemployment and Not in Education, Employment or Training (NEET) rates are problematic to the extent that they are sensitive to changes in the active population and even to seasonality. In a landmark

book, *Social Indicators: The EU and social inclusion*, Atkinson et al. (2002) consider the principles underlying the construction of policy-relevant indicators and the properties of indicators, and identify the main difficulties in the construction of comparable, robust and reliable social indicators that account for social policy outcomes.

With these caveats in mind, the literature has identified two methodological strategies for social tracking: mixed methods research and multi-sectoral integrated approaches. Mixed methods research is gaining widespread recognition as a more thorough approach for relevant evidence-building (Adato, 2008; Devereux et al., 2013). They draw on both quantitative and qualitative approaches, enabling a better understanding of outcomes and their related pathways (Devereux et al., 2013). As stated by White (2008), the use of mixed methods leads to a more policy-relevant and contextualised impact study, one that could eventually be more useful to prospective multi-sectoral interventions. The advantage of a multi-sectoral approach, as part of a more robust evaluation methodology, is its ability to elucidate the separate pathways through which outcomes are strengthened or weakened. This evidence is needed to provide insights into multi-sectoral policy design (Samson et al., 2015). If the allocation of resources stems from fragmented single-sector evaluations, more effective and long-term policy actions are less likely to be explored (Remme et al., 2014).

1.2. Budgetary mainstreaming

Budgetary mainstreaming means including cross-cutting policy goals in all phases of the policy cycle (design, preparation, implementation and evaluation) of the relevant programmes, affecting decision-making across government (Yamin, 2005). This is not an easy task, since cross-cutting policy issues and responses are ambiguous and hard to identify (Smith, 2021).

Mainstreaming is one of the meta-policy instruments of cooperation employed at EU level and its development is partly related to the failings arising from applying a sectoral approach to trans-sectoral issues (Halpern et al., 2008). Traditionally, government responses were prone to being specialised and fragmented (Bouckaert et al., 2010). This is an increasingly untenable approach for contemporary problems, which are complex, cross administrative and sectoral boundaries and require more 'joined-up' public interventions (Head and Alford, 2015).

At EU level, mainstreaming also serves to reinforce the Commission's steering capacity through the systematic adoption of specific tools and techniques, and to rationalise European public finances and policies (Halpern et al., 2008). The involvement of several actors and supportive organisational structures (Nunan et al., 2012) is crucial in order to share a common understanding of the nature of a problem and its potential solutions: this is classified as a policy image or policy frame (Mazey, 2000).

A crucial role is played by governance regimes and institutions, as well as by the involvement of well-functioning mechanisms for coordination and coherence between policy areas across government (Smith, 2021). Building on this collaboration, governments should ideally act proactively, meaning decision-making should be based on the 'logic of addressing a complex problem' (Cejudo and Michel, 2017: 158), to construct an overarching policy frame able not only to define, but even to operationalise the cross-cutting policy issues (Smith, 2021).

Mainstreaming common EU objectives in EU spending is a stated task the Commission has decided to pursue in recent years. The Commission is currently tracking expenditure linked to the climate, biodiversity, clean air and migration. More recently, the Commission has provided qualitative information on two key common objectives: gender equality and the Sustainable Development Goals (SDGs). Yet these mainstreaming exercises, as will be illustrated in more detail in Section 4, do not follow the same approach. The climate and biodiversity mainstreaming approach builds on an ex ante

identification of 'green spending' based on science-based and measurable green externalities of public investment. By contrast, the mainstreaming of gender and the SDGs adopts a rather procedural understanding of mainstreaming practices, based on iterative processes and stakeholder involvement. Building on a comparison of the two mainstreaming approaches, in Section 4 we will further investigate which of these approaches is most suited to mainstreaming the EPSR objectives.

2. OVERVIEW AND BENCHMARKING OF THE EXISTING EU SOCIAL TRACKING SYSTEMS

With tighter budgets and growing public attention being paid to the effectiveness of EU policymaking, it has become increasingly necessary to analyse the performance, impact and added value of EU-supported programmes and initiatives (Laffan and Schlosser, [2016](#); Mause, [2019](#); European Commission, [2018a](#); Zamparini and Villani-Lubelli, [2019](#)). The regulatory framework of the 2014-2020 programming period for the European Structural and Investment (ESI) Funds placed a strong emphasis on the need to evaluate the relevance, effectiveness, efficiency and coherence of the programmes supported (European Commission, [2018a](#)). For the first time, Member States were required to measure the impact and the contribution to the EU objectives of interventions co-funded by the EU structural funds, including the ESF, the main financial instrument supporting employment, ensuring fair job opportunities, supporting upskilling, and reskilling, and promoting social inclusion for over six decades (European Commission, [2018a](#)).

Compared to the previous programming period (2007-2013), the 2014-2020 Common Provisions Regulations (CPRs) introduced some mandatory steps to follow with the aim of measuring said effectiveness, such as harmonised requirements regarding ex ante evaluations and mandatory evaluation planning. The 2014-2020 CPR required at least one impact evaluation for each priority during the programming period (European Parliament and Council of the European Union, [2013](#)). With the 2021-2027 MFF, the same performance-based approach was consolidated.

The next section describes the current system in place to monitor and evaluate EU social spending. It covers the M&E system in place for the ESF, the Youth Employment Initiative (YEI) and the Fund for European Aid to the most Deprived (FEAD)). It draws attention to criticisms of the 2014-2020 M&E system and describes the key novelties of the 2021-2027 ESF+. However, it does not describe in detail the monitoring and evaluation system of the European Regional Development Fund (ERDF), which plays an important role in financing social spending in the EU budget, notably supporting investments in education (notably schools), social inclusion (notably housing) and labour market infrastructures. The box below provides a short illustration of the key indicators used in the ERDF.

Box 1: ERDF social objectives and key common output and result indicators

In line with the CPR 2021-2027, the ERDF is meant to support the thematic objective of a more social and inclusive Europe implementing the EPSR. In particular, the ERDF aims to:

- enhance the effectiveness and inclusiveness of labour markets and access to quality employment through developing social infrastructure and promoting the social economy;
- improve equal access to inclusive and quality services in education, training and lifelong learning through developing accessible infrastructure, including by fostering resilience for distance and online education and training;
- promote the socioeconomic inclusion of marginalised communities, low-income households and disadvantaged groups, including people with special needs, through integrated actions, including housing and social services;
- promote the socio-economic integration of third-country nationals, including migrants through integrated actions, including housing and social services;
- ensure equal access to health care and foster resilience of health systems, including primary care, and promote the transition from institutional to family-based and community-based care.

Contrary to the ESF and ESF+, the ERDF is targeted at infrastructure investments. To monitor the social spending under the ERDF, the CPR, as well as the fund-specific regulation, identifies 7 common output and 7

result indicators. The former measure the surface area of new or modernised facilities for employment services, the classroom capacity for new or modernised childcare, educational, temporary reception or healthcare facilities, or social housing. The latter measure the annual users of the same new or modernised facilities, services or housing. In addition to the common indicators, the ERDF include programme-specific output and result indicators.

2.1. Monitoring and Evaluation (M&E) systems in place for the ESF, YEI and FEAD

2.1.1. Monitoring

The CPR for the programming period 2014-2020 supported the strengthening of monitoring requirements, putting more emphasis on pursuing a reinforced intervention logic, the need for a result-oriented approach and a strengthened system of indicators. According to the specifications of the CPR, both Member States and Managing Authorities (MAs) are responsible for gathering and delivering both qualitative and quantitative data. They are also asked, relying on such data, to guarantee the robustness and reliability of Operational Programmes (OPs) throughout their implementation, as they progress towards their targets. In particular, Managing Authorities have to verify, in their management declarations, the reliability of performance data stated in the annual implementation report, with the additional assurance of Audit Authorities (European Court of Auditors, [2021a](#)). The Commission has the power to suspend interim payments if serious deficiencies in either the quality and reliability of the monitoring system or the data on indicators are reported. However, to date, this has never happened and could be forestalled by the provision of additional information by the Member States (European Court of Auditors, [2021a](#)).

In terms of indicators collected for the ESF, programme monitoring for the ESI Funds is built on two types of indicators: **financial indicators** and **performance indicators**. Financial indicators report on the total amount of eligible expenditure entered into the accounting system. Among performance indicators, two types are required to be collected based on the CPR: output and result indicators. The former relate to the operations supported and the latter to the priority concerned (European Parliament and Council of the European Union, [2013](#)); Article 27.3). According to the CPR, indicators should be clearly defined, and closely linked to the activity, the measurement unit selected should be specified, and they should be periodically measured (European Commission, [2014](#)). Data quality is defined following the guidelines established by Eurostat and based on six criteria: i) accuracy; ii) comparability; iii) coherence; iv) relevance; v) timelines and punctuality; vi) accessibility and clarity³. For the Directorate-General for Employment, Social Affairs and Inclusion 'it is advisable to set indicators which are as close as possible to the activities conducted under the respective investment priority' (European Court of Auditors, [2017](#)).

The monitoring systems used for the ESF, YEI and FEAD shared a common set of indicators, as provided for in the regulations (see Annex I of the ESF Regulation: European Parliament and Council of the European Union, [2013](#)). The use of Programme-Specific Indicators (PSIs) was allowed to facilitate the observation of indicators key to the intervention logic and objectives of programmes, but cannot be monitored directly using any of the common indicators. All common indicator data regarding participants entering ESF/YEI/FEAD-supported operations were initially recorded and stored following standard regulations, providing information on the given population, employment or training status at a given time. Furthermore, indicators are accompanied by standard definitions applied by Eurostat and complementary remarks.

³ <https://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf/f0fdc8d8-6a9b-48e8-a636-9a34d073410f>

All common ESF and YEI longer-term result indicators and the common output indicators on homelessness and housing exclusion and rural areas were – as a minimum requirement - to be reported on the basis of representative samples of participants. Data collection for all participants is considered as a sample covering 100% of the participants and fully representative. All common indicator data relating to participants are supposed to be broken down by binary gender categories⁴. Country-specific definitions complement the indicators covering disadvantaged participants or groups, where no EU-wide definition is available or to avoid a considerable administrative burden (European Commission, 2018a). Member States were responsible for informing the Commission if they used different definitions for less represented or disadvantaged groups. Among these groups, there are definitions for: migrants or people with a foreign background; marginalised individuals and communities such as the Roma; participants with disabilities; and the homeless or individuals affected by housing exclusion.

Specific provisions applied to YEI-supported programmes, for which the common ESF output and result indicators set out in the regulation were to be recorded in the same way as for ESF-supported programmes, except that a breakdown by category of region is not required for YEI indicators.

The FEAD M&E was put in place in line with the requirement to respect the dignity of the most deprived. Within the scope of OP II (Social inclusion of the most deprived persons), the FEAD regulations envisaged programme-specific indicators. Despite being the innovative element of the FEAD, the accompanying measures in the Member States cannot be compared due to their diversity. Consequently, the Commission did not define any common indicators on the provision of accompanying measures required for the OP I type of programmes (Food and/or basic material assistance), nor indicators measuring their results. Establishing common indicators for accompanying measures would also be contrary to the principle of proportionality to be applied under the FEAD, given that accompanying measures are limited to 5% of the support.

The European Commission (EC) Guidance fiche ‘Monitoring under FEAD’ established that in order to reduce the administrative burden, the following common indicators may be reported on the basis of informed estimates (e.g. using sampling approaches):

- Output indicators on the number of meals, the number of food packages distributed, and the share of FEAD co-financed products in the total volume of food;
- Result indicators on end recipients, which relate to migrants, participants with a foreign background, minorities, people with disabilities and homeless people; these indicators must be based on Member States’ national definitions, whereas those relating to age or gender are considered to be self-explanatory.

2.1.2. Evaluation

By collecting information on financial input, outputs, results/outcomes, and on other contextual factors, monitoring prepares the information required for a subsequent evaluation. As stated in the CPR (Article 54), Member States are instructed to provide the resources necessary for carrying out evaluations and ensuring that procedures are in place to produce and collect the data necessary for evaluations, including data related to common indicators, and where appropriate, programme-specific indicators. Evaluations can be carried out internally or externally by experts independent of the authorities responsible for implementing the programme and can be carried out ex ante, mid-term or ex post.

⁴ As of the 2021-2027 period, participants may be recorded in non-binary categories provided these are recognised by national law.

In the 2014-2020 MFF, ex ante evaluations were carried out under the responsibility of the authority responsible for the preparation of the programmes. They were submitted to the Commission at the same time as the programme, together with an executive summary. The CPR for the 2021-2027 period no longer requires mandatory ex ante evaluations. Mid-term evaluations are to focus on assessing the extent to which the programme is designed appropriately and operates in a way that is conducive to achieving its expected effects. They will look primarily at processes and implementation, and can only consider outputs and outcomes to the extent that they are realised by the time of the assessment. Ex post evaluations will be carried out by the Commission, or by Member States in close cooperation with the Commission's personnel. They are to examine effectiveness in achieving objectives, efficiency in the use of resources, relevance of the objectives vis-à-vis underlying needs, coherence with other interventions and added value compared to what could have been achieved at national level.

The overall approach to evaluations should be linked to the intervention logic and be in tune with the specific objectives and longer-term aims of an OP. It is again worth bearing in mind that transformative responses to challenges and necessities that are country-specific or regional are intrinsically long-term. Hence, they are not only influenced by the implementation of an OP, but highly likely to be affected by other external factors.

Two broad categories of impact evaluations are widely recognised and used for ESF-related initiatives which, according to a study on monitoring and evaluation, should complement each other (European Commission, [2018a](#)):

- **Counterfactual Impact Evaluation (CIE)** uses control groups, assesses how much of the change is due to the intervention and for whom, and compares the effects of different instruments (or the same instrument applied to different target groups). Rigorous quantification of the impacts of programmes involves counterfactuals which allow for the identification/estimation of what may have happened in the absence of a specific intervention. Defining such counterfactuals requires the identification of a control group (comprising people who might have been targeted, but were not subject to the intervention/programme) and comparison with a group of programme participants (the treatment group).
- **Theory-Based Evaluation (TBE)** follows each step of the intervention logic identifying mechanisms of change and answering questions about why and how an intervention works. This approach mainly produces a qualitative estimate of the impacts.

2.2. Shortcomings and limitations of the ESF M&E systems

Overall, the evaluations produced by the ECA, the Commission itself, and by other contracted organisations tend to agree on the overall positive outcomes of ESF and ESI Fund-supported initiatives, and that indicators are on track to achieve their targets and follow a robust intervention logic.

However, both shortcomings and limitations were identified during the interim evaluation and studies produced on ESF-implemented programmes, specifically on the quality of data available and the indicators used to provide evidence for the successful implementation of OPs (ECA, [2018](#)). Many of these concerns have been addressed in the current MFF 2021-2027 in the ESF+ governance, as detailed in the next sections. For completeness, the main criticisms are nevertheless reported here to provide the European Parliament with a basis for an independent assessment of the key novelties introduced for ESF+.

To start with, qualitative surveys with Member States, and the information available on the fulfilment of ex ante conditionalities⁵, indicate that for many Member States it had been challenging to implement appropriate monitoring systems, in some cases leading to substantial delays in the implementation of the programme.

MAs surveyed for an external study on the M&E systems of the ESF (European Commission, [2018b](#)) reported some significant difficulties in setting up monitoring systems for the 2014-2020 programming period, due to challenges in understanding the reporting requirements and/or in coordinating with beneficiaries to establish procedures to ensure that the requirements were fulfilled. Some monitoring committee members also noted that the complexity of the monitoring requirements was a challenge, that the regulations were too detailed, and that the reporting requirements for the Annual Indicator Report Series (AIRS) were complicated and time-consuming.

The complexity of some common indicators was another challenge mentioned by the MAs surveyed in the evaluations carried out to assess the performance of the M&E systems. The use of complex common indicators, for instance combining common output and longer-term result indicators⁶, to describe programme beneficiaries increased the demands on data collection and processing. This inevitably adds to the already high associated administrative burden.

Some indicators have been criticised as misleading, specifically those describing social exclusion and disadvantage. Output indicators for participants may not necessarily reflect disadvantage. For the programming period 2014-2020, targets for common result indicators were set in relation to common output indicators. However, this limited the scope for capturing the specific target groups of a given programme, thus provoking the proliferation, in some cases, of programme specific indicators.

The use of national definitions for certain common output indicators referring to disability, migrant, or minority status, or general disadvantage was allowed either because no common definitions were available at EU level, or to avoid the administrative burden of imposing harmonised definitions that would conflict with national-specific definitions. Furthermore, the comparability of these indicators is potentially compromised by under-reporting. This could be because of data protection regulations, even though the regulations enable data collection for these operations. Comparability also concerns the unit of measurement. In the CPR, there is no specific provision for the units to be used for reporting on result indicators, leaving MAs free to use either percentages or absolute numbers (European Commission, [2018c](#)).

Weaknesses in counting procedures for participants were also identified as a shortcoming of the M&E system (European Commission, [2018b](#)), and these persist in the current 2021-2027 period. Participants returning to an operation should be counted only once, as established by the reporting requirements. Data for common output indicators, for instance, should describe the number of individuals who benefit from support, with just one result for each person, not the number of participations. Double-counting or failure to cap the value for the indicator 'population served' at the number of residents in the area can result in inconsistent values and provoked several issues for MAs. Examples included: the administrative burden created by the revision of data; significant alterations of the success rate of an operation; difficulties in linking expenditure with outputs and results; and inconsistencies with other procedures of data collection and statistics at EU level, which are based on participations.

⁵ A set of minimum strategic and regulatory requirements was introduced in the 2014-2020 period, which Member States had to meet at the beginning of the programme implementation phase and before they could receive any funding under the ESI Funds. They were designed in order to better address, in the early stages, any problems limiting the potential of Member States and regions to attain the intended objectives.

⁶ E.g. combining an output indicator such as 'above 54 years of age who are unemployed, including long-term unemployed, or inactive not in education or training' with the common longer-term result indicator 'disadvantaged participants in employment, including self-employment, six months after leaving'.

The drawing-up of representative samples frequently created difficulties for MAs and stakeholders, thereby complicating the CPR requirements to provide reliable data. The Commission, relying on advice from the Centre for Research on Impact Evaluation (CRIE), issued minimum threshold requirements for drawing such samples. It also required the sampling design, method of sample selection and actual sampling methods to be developed in advance of any data collection and to be fully documented at all stages of the process. However, these procedures were found to be difficult to achieve, and liable to aggravate financial and administrative burdens.

Difficulties also arose in collecting data directly from participants, according to survey responses taken from evaluations of ESF programmes. The biggest concern is the quality and volume of data requested. Member States and beneficiaries have flagged difficulties in providing personal data for several reasons, mostly related to the strictness of data protection laws, sometimes conflicting with data requirements, or individual reluctance to provide personal, sensitive information. In particular, participants are often hesitant to share sensitive information or data involving third parties, e.g. household indicators or data on minors, which require parental consent.

According to a Commission Staff Working Document (European Commission, [2020a](#)), data for young people are inherently more difficult to collect, especially for minors, NEETs, and young people with a disadvantaged background or coming from rural areas. These issues make reporting for these categories of participants harder, and monitoring and evaluations of programmes for these groups are likely to be affected by under-reporting, either given the difficulty in gathering this kind of information systematically or because the participants do not want to disclose such information. Moreover, Member States have varying interpretations of disadvantage, e.g. regions or countries hit by economic hardship may consider a bigger share of the population as disadvantaged.

Another shortcoming is the misalignment of timelines between the Member States' annual implementation reports and their respective management declarations and annual control reports (ECA, [2021a](#)). The ECA's audit of 2018 sought to establish whether the selection of projects and monitoring systems for ESF projects in the 2014-2020 programming period were results-oriented (ECA, [2018](#)). The findings from the audit showed that the information generated on the basis of the data collected was mostly outputs-oriented, that the selection of projects under the ESF was not fully results-oriented, and that the monitoring in place was still more outputs-oriented.

Regarding results, ESF programme-specific indicators measured different things in different ways. In fact, the concept of a 'result' is interpreted differently in the fund-specific regulations. For ESF-related initiatives, the ECA identified a risk that a meaningful aggregation of performance data (in particular regarding results) may not be feasible at EU level, thereby highlighting the limits of performance information usage for decision-making and corrective actions (ECA, [2019](#)). Another shortcoming linked to result indicators is related to the timeliness of information: it takes time before results are released, which again hampers decision-making efficacy (ECA, [2019](#)).

2.3. Monitoring and Evaluation (M&E) in the ESF+: key novelties

Coming into force on 1 July 2021, the ESF+ integrated the former ESF, the YEI and the FEAD, to maximise the efficiency and coherence of EU funding under a single overarching initiative. In the same spirit as the ESF, the ESF+ is the main EU budgetary instrument for improving social cohesion and workers' mobility and increasing competitiveness across Europe for the 2021-2027 period. It is also charged with supporting the implementation of the EPSR, the EU's compass for responding to social challenges across Europe (Lecerf, [2019](#)). The ESF+ operates under shared management with the Member States, meaning that the ESF+ MAs in each country allocate the money to projects that are run

by a range of public and private organisations, thereby responding to country- and region-specific needs.

As with the 2014-2020 programmes, lessons have been drawn from the experience of the previous funding period. Much of the M&E framework for the ESF+ will remain the same as for the ESF. However, some notable changes were introduced. Emphasis was placed on the need for a stronger alignment with the recommendations and country analyses provided under the European Semester, as well as with the EPSR (Lecerf, [2019](#)).

The performance of the shared management strand of the ESF+ will be measured through common and programme-specific output and result indicators (see Annex II of the ESF+ Regulation: European Parliament and Council of the European Union, [2021a](#)). Success will continue to be assessed through the comparison of the achievements with the targets (achievement ratio) as well as the comparison of the results with the outputs (success rate). Targets for outputs and results will continue to be set by Member States for each programme.

In formulating the ESF+, considerable emphasis was placed on reducing the burden of data collection and processing. Most importantly, Member State obligations for monitoring and reporting were streamlined, and the number of indicators was reduced. Common output indicators were reduced to the most relevant target groups of the ESF+ (e.g. number of inactive, unemployed, NEET, young and elderly, and low-skilled participants) and allow the calculation of the planned achievements of the fund at EU level (Lecerf, [2019](#)). Sensitive indicators with little relevance or complex definitions, e.g. 'other disadvantaged' have been dropped. Sensitive data may be collected through informed estimates, thereby avoiding the need to put difficult questions to difficult-to-reach participants. This was intended to reduce the deterrent effect of having to answer sensitive questions and to simplify data collection. Where possible, definitions for common indicators were revised.

Common indicators that are not applicable from the outset (e.g. employment in the case of pupils in primary schools) are no longer to be collected. The number of common result indicators was also streamlined, notably by merging ESF and YEI indicators. Evaluation remains the shared responsibility of the Member States and the Commission. The Commission may, at its own initiative, carry out further evaluations of the ESF+. The previous requirements for impact evaluations were maintained and the reporting of evaluation findings will be more structured in order to facilitate the meta-analysis of the available evidence, thereby facilitating the preparation of the next MFF as well as the mid-term review of the programmes. Impact evaluations of all priorities of all programmes are required, where possible, for the assessment and quantification of the impacts of the fund, as was the case for the ESF. Reporting should be more structured, making it easier to summarise and disseminate results by using, for instance, the evaluation criteria as defined in the Better Regulation Guidelines.

An enabling clause to foster a wider use of available administrative registers for monitoring and evaluation was introduced. This possibility was left to the discretion of Member States even though the Commission is pushing for increased use of administrative registers. As argued below in more detail in the recommendations, administrative data have several advantages, but also disadvantages in terms of definitions as well as practices which change significantly across Member States. The administrative registers in certain countries are devolved to sub-national level and can be more difficult to collect. Some MAs do not even know where to ask in certain regions or if these data really exist. Data protection restrictions can be a further problem in this respect. This requires time, but it is important for the Commission to push in this direction and invest significant resources in exploring the use of these data for policy evaluation, drawing on the work of the Competence Centre on Microeconomic Evaluation (CC-ME) active at the Joint Research Centre (JRC) since 2016.

Support for the quantification of net effects of results through counterfactual impact evaluations will continue. As counterfactual evaluation is heavily reliant on data on comparable non-participants, MA and contractor access to administrative registers should be facilitated while complying with data protection rules. MAs may investigate what data registers could be used for M&E purposes and to secure access to them from the outset. The use of macroeconomic models for quantifying high-level impacts, e.g. on the employment rate, is being improved by structured data about ESF+ investments at NUTS 2-level⁷.

2.4. Monitoring and Evaluation (M&E) in the RRF

The establishment of the RRF as a strategic response to the COVID-19 crisis formally introduced the adoption of a performance-based approach to EU spending. The main novelty of the RRF is in the design of National Recovery and Resilience Plans (NRRPs), which are performance-based contracts agreed ex ante between Member States and the Commission and translated into implementing decisions adopted by the Council. To monitor the implementation of the plans, composed of reforms and investments, the Commission assesses Member States' fulfilment of milestones and targets (M&Ts). These are result-driven implementation steps which monitor, in a granular way, the progress towards the achievement of a reform or an investment.

While both measure the progress of implementation, milestones are qualitative implementation steps and targets are quantitative implementation steps. The M&Ts reflect performance under each reform and investment, and thus the results achieved, in order to unlock payments. Member States sign an Operational Arrangement with the Commission further specifying the M&Ts, submit payment requests and receive disbursements based on a positive assessment by the Commission of the satisfactory fulfilment of the M&Ts linked to the instalment concerned. Member States implement their national plans by fulfilling M&Ts.

In December 2021, an RRF Delegated Act (European Commission, 2021b) introduced a Recovery and Resilience Scoreboard, with a dedicated website including information not only on the achievement of M&Ts, but also on common indicators created by that same delegated act. The common indicators displayed on the Recovery and Resilience Scoreboard aim to monitor the progress of the implementation of NRRPs towards common objectives and the overall performance of the RRF, and are based on data provided by national authorities. Of the 14 common indicators established by the RRF Delegated Act, 5 relate to employment and social areas:

- number of participants in education or training;
- number of people in employment or engaged in job searching activities;
- capacity of new modernised health care facilities;
- classroom capacity of new or modernised childcare and education facilities;
- number of young people aged 15-29 years receiving support.

Data on such indicators allow the Commission to monitor the progress towards the general and specific objectives specified in Article 4 of Regulation (EU) 2021/241 (European Parliament & Council of the European Union, 2021b). Each measure in the NRRP can contribute to several common indicators. Data on these common indicators are collected by Member States, which must report on them twice a year.

⁷ **Nomenclature of territorial units for statistics** (from the French version *Nomenclature des Unités territoriales statistiques*) is a geographical nomenclature subdividing the EU's territory into [regions at three different levels](https://ec.europa.eu/eurostat/web/nuts/principles-and-characteristics) (NUTS 1, 2 and 3). <https://ec.europa.eu/eurostat/web/nuts/principles-and-characteristics>

Reporting on common indicators covers the entire period of RRF implementation. In order to ensure the comparability of common indicators at EU level, each Member State is asked to adhere to the guidance provided by the Commission and can deepen the methodology for each indicator, so as to develop a more in-depth calculation methodology that is consistent at national level. For each common indicator whose computation requires estimates or specific methodological approaches that go beyond the guidance provided at EU level, Member States are required to share the methodology with the Commission. Given the short timeframe for the implementation of the plans and the political will to ensure that the measures progress as rapidly as possible, the indicators focus on outputs rather than outcomes and impact, which require more time to be evaluated.

In addition to the Recovery and Resilience Scoreboard, in September 2021 the Commission also published a Delegated Regulation (EU) 2021/2105 (European Commission, [2021c](#)) which defines a methodology for reporting on social expenditure, supplementing Regulation (EU) 2021/241 (European Parliament & Council of the European Union, [2021b](#)). The aim of the methodology is to provide, in a transparent and accountable manner, synthetic information on the social expenditure under the RRF. The methodology consists of two steps.

First, each reform or investment included in NRRPs with a social dimension should be associated with one of the following four social categories, in consultation with the Member State concerned: employment and skills; education and childcare; health and long-term care, and social policies⁸. According to the Delegated Regulation (EU) 2021/2105 (European Commission, [2021c](#)), *'reforms and investments with a primary social dimension shall be associated to one of the nine social policy areas set out in the Annex. Each social policy area shall be linked to a broader social category. A reform or investment may only be associated to one social policy area, and therefore to one social category'*. In that regard, the regulation mentions the social objectives of the Facility, i.e. its contribution *'to fighting poverty, inequalities and tackling unemployment, creation of high-quality and stable jobs, improved health care capacity, and better policies for the next generation including on education and training.'*

Second, the Commission, in consultation with the Member State concerned, flags each measure that includes a focus on children and young people, as well as on gender equality, allowing for specific reporting on expenditure.

2.4.1. The limits of M&E of the RRF

In terms of M&E, the key novelty of the RRF compared to the ESI Funds is the new M&T approach. As stressed above, Member States submit payment requests and receive disbursements based on a positive assessment by the Commission of the satisfactory fulfilment of the M&Ts linked to the instalment concerned. The introduction of this 'financing not linked to costs' approach has been welcomed as a positive step towards a new performance budgeting, a shift away – in the EU budget – from managing (i.e. 'how much have we spent?'), to the achievement of policy objectives (i.e. 'what have we accomplished with our money?' (Darvas, [2022](#)).

The assessment criteria for the NRRPs defined in the Annex to the RRF Regulation (European Parliament & Council of the European Union, [2021b](#)) make it clear that, to ensure effective monitoring and implementation of these plans, the definition of M&Ts should be clear and realistic and the related indicators relevant, acceptable and robust. Since fulfilment (or not) of M&Ts is the sole criterion to justify the disbursement (or not) of the RRF tranche, the Commission Guidance to Member States on the RRF (European Commission, [2021j](#)) explicitly invites Member States to propose M&Ts, which remain within their control and are not conditional on external factors such as the macroeconomic outlook or

⁸ The four main policy areas contain nine sub-categories detailed in the Annex of the Delegated Act (European Commission, [2021c](#)).

the evolution of the labour market. Partly for this reason, the indicators for M&Ts included in the RRF plans are nearly all output rather than outcome indicators. The focus on immediate outputs rather than the impact on the long-term consequences of the interventions allows for M&Ts to be used to measure the performance of the interventions, and thus to monitor the results of the RRF over time, while avoiding indicators that may be burdensome or difficult to manage.

Despite these positive steps towards a new performance-based approach, some limits can be highlighted that warrant reflection. The first concerns the rigidity of the current system of M&Ts, which have to be set *ex ante*. This should be done, as stressed above, in a clear and realistic way. Due to inherent uncertainty and rapidly changing conditions in policymaking, however, this is no easy task. Modern investment practice therefore often relies on smaller milestones and regular joint review to allow for revision, in line with an experimentalist logic (Gilson et al., 2008). This is not the case for the RRF which does not envisage a revision of the M&Ts. Bokhorst and Corti (2023) point here to a number of risks and challenges that this centralisation and ‘contractualisation’ may have when it comes to implementation.

First, the plans were typically formulated at unusually short notice and in a centralised manner. To speed up the processes, the M&Ts were in some cases only quickly discussed within the relevant line ministries before being agreed upon and negotiated with the Commission by the ministries of finance, mostly with very limited consultation. Second, even though Member States were expected to provide a summary of the input of, *inter alia*, local and regional authorities, and to explain how this input was taken into account (European Parliament & Council of the European Union, 2021b; Article 18(4)(q)), formal involvement is not in the RRF Regulation.

The need to define a clear and transparent set of M&Ts *ex ante*, in a centralised way and within a very constrained timeframe, *de facto* leaves little space for context-specific interpretation and adaptation. Since the governing principle in the RRF implementation is ‘comply or don’t get paid’, Member States might be tempted to commit to *guaranteed* targets and milestones, also including in the plans, reforms and investments already planned, to be sure to receive the Commission disbursement (Corti et al., 2022a). Especially for investments involving sub-national authorities, they might prefer to commit to aggregated M&Ts to make sure that in the case of non-performing territories, this would not undermine the disbursement of payments (Corti et al., 2022b).

The selection of ‘guaranteed’, overly aggregated or even vague M&Ts, however, might limit the utility of the RRF monitoring system. A Member State could be compliant with the indicators agreed with the Commission, and thus will successfully fulfil the M&Ts specified in the Council Implementing Decision (CID) and thus will receive the Commission disbursement. Yet this information will reveal little about the actual effectiveness of the measures implemented *vis-à-vis* their objectives. To be clear, the M&T arrangement in place would work formally, but would fall short in meeting the expectations of a paradigmatic shift towards performance budgeting.

3. EXPLORING ALTERNATIVE APPROACHES

The M&E system of the ESF+, described above, builds extensively on the methodological framework of the EU's Better Regulation agenda and applies several concepts derived from different evaluation traditions. However, a clear shortcoming of this M&E framework is the lack of sufficiently clear and reliable identification of impacts generated through EU social expenditure. The difficulties of finding indicators capable of showing a causal link between a specific intervention and impacts have already been noted. Moreover, as expert interviews conducted for this study amply demonstrate, making such links is especially challenging in the domain of social policy.

The previous section on the current EU system summarised some of the main shortcomings and challenges of tracking social expenditure. Despite some of the difficulties inherent in measuring impacts, such as a plethora of different endogenous factors contributing to an outcome, past experience can provide insights into some of the most crucial elements to take note of when designing social tracking methodologies. One important observation relates to the importance of common definitions and the proportionality of monitoring requirements vis-à-vis analytical needs for measuring programme performance.

With 27 different national implementation systems and a very large number of Managing Authorities and Monitoring Committees in charge of designing and implementing OP-level data collection activities, a common understanding of the type of information to be gathered is critical, albeit often lacking. The absence of such an understanding – for instance in relevant legislative texts⁹ and audit practices, or due to divergent national definitions of common indicators (ECA, [2021b](#)) – calls the comparability of data into question. It renders EU-level aggregation nearly impossible, and the low comparability severely limits analytical efforts to assess the effectiveness of the respective programmes. Aggregation is equally problematic for the large number of programme-specific indicators, the collection of which also leads to a considerable administrative burden (ECA, [2017](#)).

This section explores an alternative approach that has the potential to correct, or at least diminish, these shortcomings, derived from TBE, an approach already to some extent a feature of the current M&E framework. This study contends that more use can be made of TBE principles in monitoring (as opposed to evaluating) the impacts of social expenditure, rather than providing an entirely new conceptual foundation.

3.1. Theory-based evaluation (TBE) and its application in the current M&E framework

Before delving into the detail, it is worth looking at the theoretical underpinnings of TBE, as well as its limitations. TBE calls for having explicit theories of change about the rationale, causal mechanisms, interrelations and objectives of different forms of policy interventions. The approach is well suited to circumstances in which what would happen in the absence of the policy (the 'counterfactual') cannot easily be ascertained. A succinct definition published by HM Treasury (HM Treasury, [2020a](#)) states: 'theory-based evaluation is explicitly concerned with both the extent of the change and why the change occurs. In addition, it often considers the context at the same time that the intervention is being implemented'.

In short, TBE relies on having a coherent understanding of what it is that the policy is expected to achieve and how it will work in the context in which it is applied. In a sense, the approach reverses the

⁹ Or in the case of 'results', between the ESF and ERDF Regulations, see ECA ([2018](#)). The issue raised by the ECA has been addressed for the 2021-2027 period.

conventional way of looking at a policy by seeking to define at the outset what long-term impact it is intended to achieve and the problems to be solved. Working backwards, the direct outputs of the policy will contribute to these impacts, but the connections need to be specified. Then, the outputs have to be mapped to the inputs provided (financial resources and the contributions of personnel) and the activities they generate.

The usual starting point is to analyse the context in which the policy is applied, then develop policy interventions intended to alter the different variables in question in the desired direction. Doing so relies on having a theory of how the policy intervention (logic of intervention) will lead to change which, in turn, calls for an understanding of the different stages of a transformation. In the TBE terminology, even if impressive outputs can be demonstrated, they are ultimately only of value if they subsequently translate into the impacts the policy was intended to generate. Spending on a policy, the input, can be judged in at least three ways:

- It generates direct outputs, such as more training places or better quality social care;
- These direct outputs lead to outcomes for target groups; examples could be higher employment or reduced poverty;
- The outcomes, in turn, affect societal impact or 'utility', for example greater equality or enhanced inclusion and thus the attainment of the priorities set out in the EPSR or similar initiatives.

When applied to the social tracking of budgetary spending, an obvious complication is that the policy will have been the result of many compromises and political deals, making identification of the variables to track more problematic. Moreover, the social sphere has multiple stakeholders, both in the delivery and receipt of policies, and the amounts spent by the EU are often small and may be narrowly targeted. Substantive and welcome impacts may arise, but it will frequently be far from clear which of the multiple policy interventions was most consequential and whether a contribution to an aggregate policy effort made a visible difference. As stressed in more than one of the interviews conducted for this study: correlation, in short, is not causality.

Nevertheless, a disciplined framework can be valuable in drawing attention to the variables and metrics to be used to monitor and evaluate the policy. Impact is intrinsically longer-term, and it may not always be easy to track sufficiently rapidly to be useful for immediate policy purposes, such as changing the focus of expenditure programmes. To take one (somewhat extreme) example: the early intervention prescribed in social investment strategies to boost a child's opportunities in adulthood will only be shown to have worked when the child has left formal education. It may be possible twenty years later to detect a qualitative change, but attributing it to expenditure in year $t-20$ is unrealistic. Yet some valuable information for policy purposes may well be obtained from looking at qualitative information about directions of change.

TBE is already meant to be used to assess cohesion policy interventions. For instance, the common indicator toolbox for ESF+ in the 2021-2027 programming period states that a 'clear intervention logic constitutes the essential prerequisite for the assessment of the programme proposals' (ESF Data Support Centre (2021), p. 7). An intervention logic helps to establish whether a policy's outcomes flow from the policy itself or are the result of exogenous factors. Key questions include whether the logic was appropriate for achieving the goals of objectives as ambitious and varied as those in the EPSR.

The second key component is 'achievements': what matters is whether the policies contribute to the results its architects sought and anticipated. Defining achievements in this way requires elaboration, especially when cumulative effects are of interest. They may include progress towards meeting the immediate targets of policies, which can be understood in the standard evaluation sense of 'effectiveness'. But the notion of achievement needs to encompass impacts not easily captured by

indicators. For this purpose, it may be useful to explore whether milestones – already part of the monitoring framework for the RRF – can be built upon to capture the stages in the intervention logic leading towards the outcomes and impacts sought by policymakers.

The need to strike the right balance between conceptual simplicity and complexity is paramount. The boundary between what is attributed (success or failure) to the theory (was it well-founded?) and the implementation (were there inadequacies that undermined a valid theory, or vice versa?) is tricky. Having a well-conceived theory of change is vital. If the theory is defective because it does not provide a direct understanding of how the intervention is supposed to work, evaluation or monitoring will be more difficult.

3.2. Advantages and drawbacks of TBE

While extensively used in evaluations, TBE has to be undertaken with care because it is neither a panacea nor necessarily straightforward. This section points to several features of the approach, as well as potential pitfalls.

There are several components of TBE, requiring solutions to the following:

- Analysis of why an intervention is needed, its likely incidence on target groups and the impact it is intended to achieve;
- Clarifying how the proposed policy will interact with other initiatives or resources aimed at broadly the same objectives, including initiatives that are implemented at EU, national, regional or even local level;
- Explaining what aspect of the problem the intended policy will target and how it will contribute to outputs, outcomes and, in due course, impacts;
- Showing why the intervention will facilitate this transition, ideally distinguishing between what the policy can expect to achieve and how it is implemented.

Different theories of change might be applicable at different stages or for different stakeholders, posing the question of whether credible theories can be developed to reflect the European Parliament's particular perspective in tracking social expenditure. It is also essential to look at how crucial, internally consistent and credible the assumptions behind the theory are. Another question is whether the theory can shed light on how useful it is to find data on different stages of the policy.

A programme can be thought of as having a number of aims and features:

- Its primary rationale should be to respond to an identifiable need as part of the priorities of the EU;
- It should be clear who is intended to benefit from it and how target groups are distinguished from other groups;
- The expected respective contributions of different stakeholders and levels of government should, as far as possible, be differentiated, enabling a more precise assessment of the specific EU intervention; ultimately it is about true social impact.

In addition to problems of comparability, and despite the large number of indicators and a complex M&E system, the effectiveness of several areas of social policy, even if linked to EPSR principles, is difficult to assess, as explained in recent ECA work¹⁰. Because of the need for proportionality

¹⁰ Long-term unemployment (ECA, [2021b](#)) and especially child poverty (ECA, [2020a](#)) are two examples cited previously.

highlighted above, it is not possible to provide detailed monitoring data on all aspects of EU policy. The choice of the areas of focus, therefore, depends on considerations for measuring specific aspects of programme performance and prioritisation.

3.3. A blind spot: the 'black box' approach to causal pathways

The common indicator toolbox referred to above also notes that the identification of programme-specific indicators should be based on an intervention logic. Nonetheless, the requirement to collect impact indicators systematically is dismissed. This is because 'given the difficulties in collecting sound and timely data for impact indicators and in proving a causal link, their use for monitoring purposes is not recommended' (ESF Data Support Centre (2021), p. 9).

The assessment of impacts is thus entirely relegated to the realm of programme evaluations, which occur periodically, depending on legislative requirements specific to each intervention, and (except where an interim evaluation is required) usually *ex post*¹¹. What is missing is a potential bridge between monitoring and the questions typically posed in evaluations. As per the EU Better Regulation recommendations, the analysis of overall (and sometimes specific) intervention logic is an integral part of these exercises. However, they usually form part of the assessment of the relevance of expenditure programmes (i.e. the relationship between evolving needs and stated objectives), but constitute only a marginal part of the analysis of their effectiveness (i.e. success in achieving objectives). This practice – though not universal and not unique to evaluations of cohesion policy – focuses on observed outputs and assumes underlying causal processes rather than confirms them.

As noted in Section 2, a tool deployed to redress the apparent lack of reliable evidence on the impacts generated through social interventions is counterfactual impact assessments (CIAs). Heavily encouraged during the 2014-2020 programming period and further mainstreamed under the 2021-2027 MFF, CIAs circumvent the thorny issue of causality by attributing the difference between an assumed 'counterfactual' scenario, i.e. absence of the intervention, and observed impacts to the policy programme in question. Combining the TBE approach with the increasing use of CIAs to assess ESF operations undeniably yields valuable evidence and input for a more evidence-based formulation of policies. Nonetheless, it still falls short of a systematic assessment of the different mechanisms resulting in observed impacts. This creates several problems:

- A lack of detailed understanding of how, and through which specific mechanisms, public investment contributes to the achievement of higher-level policy objectives (in this case, principles of the EPSR);
- The actual extent of the contribution. With several confounding variables at play, this might be under- or overstated;
- Insufficient evidence to fine-tune future interventions in view of broader objectives.

It should be noted that, as shown by the documentation publicly available and the stakeholder input collected for this study, policymakers are well aware of these limitations. In fact, the changes introduced in the 2021-2027 programming period were an attempt to optimise the use of available evidence, while also keeping in mind the need to keep the administrative burden associated with monitoring at a manageable level.

¹¹ While mid-term evaluations also aim to draw conclusions on impacts, they can only assess 'expected' impacts due to the early stage of implementation at the time they are conducted.

3.4. Potential methods to measure impacts and policy utility: contribution analysis and contribution tracing

There are several empirical approaches used in evaluations adopting a TBE approach, but they are prone to being very demanding in terms of data requirements. For the purposes of social tracking – understood to be about monitoring that is potentially aimed at correcting a programme that is not performing as anticipated, as opposed to ex post evaluation – some simplification is essential. Keeping in mind the inherent difficulties, this section explores two alternative approaches, namely: contribution analysis and contribution tracing.

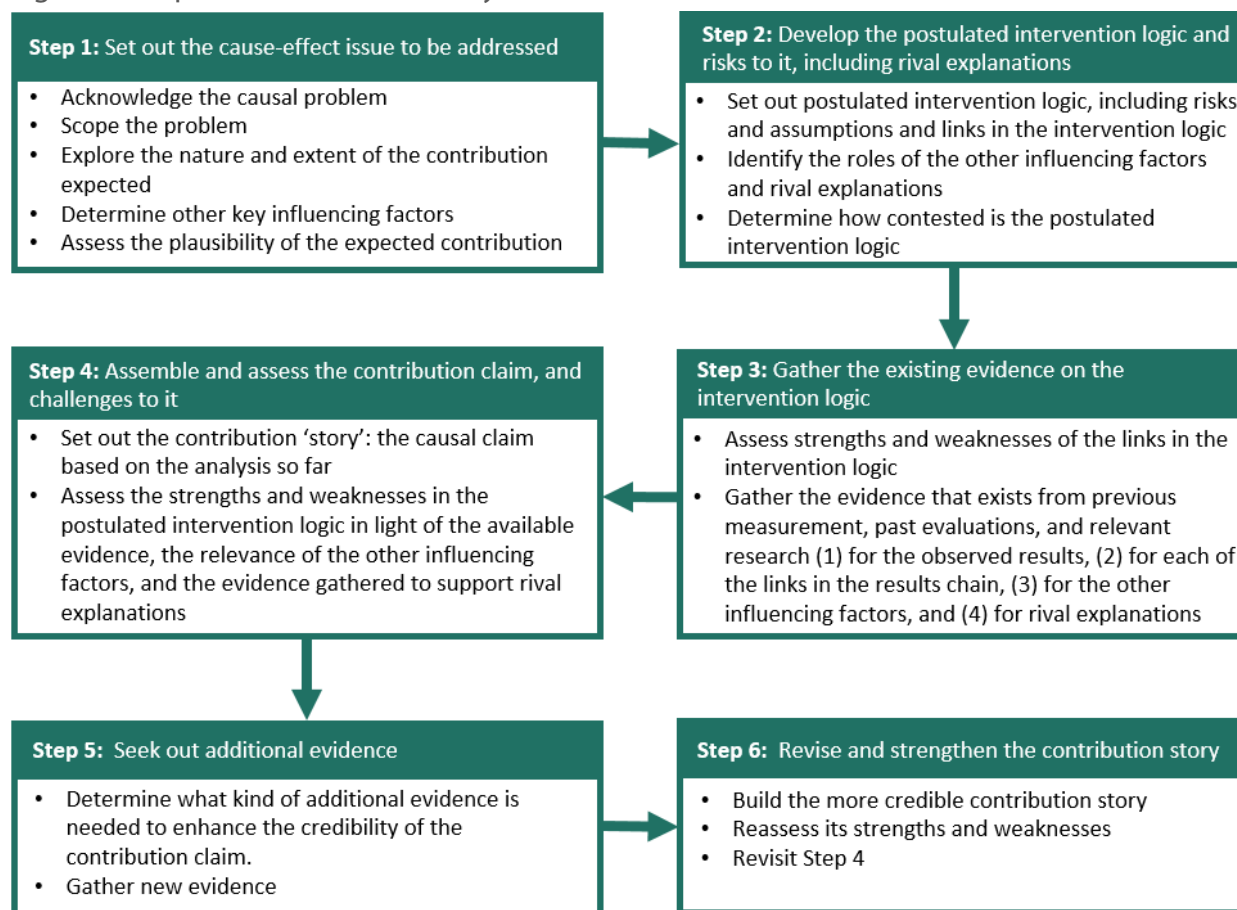
3.4.1. Overview of contribution analysis and contribution tracing

Contribution Analysis (CA) is a TBE approach which seeks to make credible causal claims about the results of interventions. CA is used to understand the likelihood that the intervention has contributed to an observed outcome. This is especially helpful to address the issue of attributing results to a specific intervention where there are other influencing factors at play (HM Treasury, [2020b](#); Mayne, [2012](#)). In the context of social Europe, where so many actors and stakeholders are involved in the different stages of the policy process, distinguishing different contributions is vital.

CA assumes that specific interventions are unlikely to be the sole causes of change. Instead, they are treated as contributing causes and CA therefore examines ‘contribution claims’. The approach attempts to examine the intervention logic, or ‘theory of change’ against the data available on observed results and the assumptions it builds on, as well as other influencing factors (Befani and Mayne, [2014](#)). The step-by-step approach (see figure below) makes use of a wide range and type of evidence to assess the credibility of these contribution claims.

CA does not provide estimates of the size of the contribution and is, in essence, qualitative. It is therefore not suitable for comparing interventions in separate policy areas. However, it helps to infer causality in real-life interventions and thus reduces uncertainty about how impacts were achieved. Social tracking that at least allows the direction of change to be established may be valuable where accurate quantitative data are hard to find.

Figure 2: Steps in Contribution Analysis



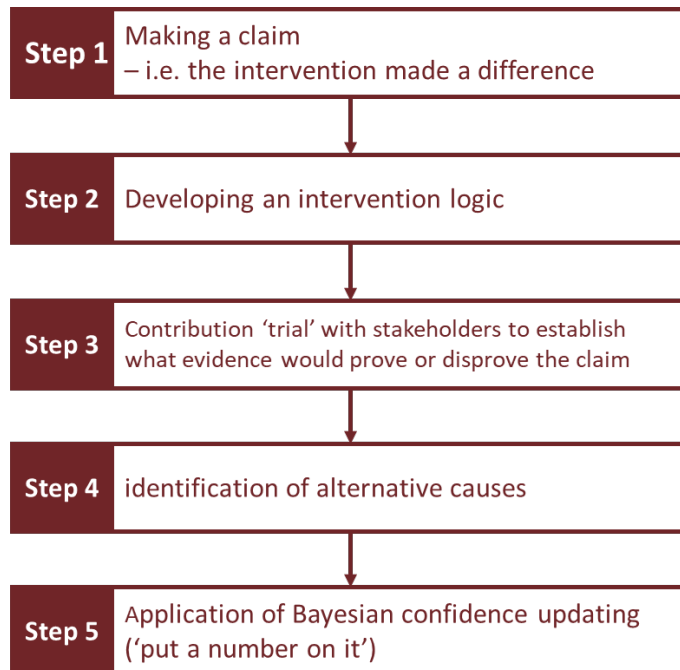
Source: Own elaboration based on Mayne (2012)

The second approach reviewed here builds largely on CA, and complements it with elements borrowed from process tracing and Bayesian updating. Contribution Tracing (CT) is a mixed qualitative and quantitative method that, similarly to CA, shifts the focus from assessing impact to assessing *confidence* about the impact. CT, following the logic described for CA, gathers evidence supporting or disproving contribution claims, making use of an intervention logic/theory of change. The evidence is then analysed using mathematical formulae building on the Bayes theorem. In practice, it updates the confidence in the likelihood of a contribution claim conditional on finding new evidence. The estimate is based on three pieces of information:

- The probability that the contribution claim is true before observing (finding) new evidence;
- The probability of finding evidence if the contribution claim is true;
- The probability of finding the same evidence if the contribution claim is false.

The steps of CT are summarised in the figure below (Bamanyaki and Holvoet, 2016; Mayne, 2012; HM Treasury, 2020b).

Figure 3: Steps in Contribution Tracing



Source: Own elaboration based on Mayne (2012) and HM Treasury (2020b)

CT as an approach has several advantages. First, it is participatory and minimises confirmation bias through the involvement of stakeholders during the assessment of relevant evidence to prove or disprove claims. It also provides a rigorous assessment of the evidence available to increase certainty on whether a specific policy intervention achieved the desired results/impacts. It also has shortcomings, most of which it shares with CA. It cannot be used to compare different programmes and can only be used with a certain time lag in order for the evidence – ‘traces’ – to become visible (HM Treasury, 2020b).

3.5. Practical applicability

Any conceptual framework to be used in the complex setting of monitoring and, subsequently, evaluating EU social policy will have to take into account several factors. This section will consider the two above approaches along four main criteria:

1. Clarity of information generated for policy purposes;
2. Data requirements and expertise needed for implementation;
3. Plausibility of use for regular monitoring as opposed to one-off evaluation exercises;
4. Potential for simplification, not least to be viable in a political context.

The overarching question that this section aims to answer is whether these approaches, which were developed for evaluation, could be adapted (or simplified) for monitoring and tracking. If so, what changes would have to be introduced to make their practical application feasible? A possible parallel is with monitoring of climate change policies, as explained by McConnell (2019), who illustrates the approach for a five-year programme on adaptation to climate change, involving mainstreaming into policies. Suggestions for a TBE approach to monitoring in water management are offered by Hermans et al. (2012), but involve a more complex analysis that is unlikely to be applicable on a regular basis by the EP.

McConnell (2019) notes that theory-based monitoring (TBM) will not always yield the desired results, but [p. 222] nevertheless asserts that it can 'be helpful in providing a basis from which to viably infer impact and reduce uncertainty about the contribution being made by the programme'. As he describes it:

TBM (as opposed to TBE) 'relates to the on-going and systematic collection of data that pertain to specific mechanisms within a causal chain to ascertain change and progress towards achieving a desired result'.

Indicators can provide useful insights into the effects of policies, but are insufficient on their own. Instead, the approach calls for information on processes that underpin policy achievements. The process, in this context, encompasses the steps in a theory that need to be followed to arrive at desired outcomes and impacts, as well as obstacles to success or unintended consequences as the programme unfolds. Where multiple stakeholders are involved, it should make some allowance for different 'theories' behind their respective approaches and expectations. For the EP, this suggests that there should be an attempt to ascertain whether its own specific model can be identified and what distinguishes it from other stakeholders.

A further point emphasised by McConnell is that the pathways from inputs to outcomes may not be linear and it is vital to collect evidence on whether expectations are realised or need to adapt. This challenge is more pronounced in complex systems where changes in behaviour or attitudes towards social (or societal) goals may evolve. Identifying what can be thought of as the 'moment of change' will be useful.

Different outputs may contribute to overall programme aims, and need to be integrated in a coherent manner, following the principles of contribution analysis or tracing described above. The quality of monitoring can be enhanced by attempting to define intermediate outcomes (achievements beyond the direct outputs of an intervention), even if not they do not fully capture eventual desired impacts. In this way, milestones crucial to achieving impact can be put in place. However, beyond monitoring whether or not milestones are attained are the questions of how and why, shedding light on the validity of assumptions about how a policy will function.

The current M&E framework is also built around the understanding that TBE ought to be extensively applied in assessing the successfulness of social expenditure from the EU budget. With this in mind, specific arrangements and conceptual frameworks should already have been set up at the policy design stage. The extensive use of intervention logic analysis – the very foundation of both the CA and CT approaches – makes the tailoring of the framework at least theoretically possible. However, evidence collected for this study suggests that the approach is not often applied as rigorously as might be expected.

Milestones, targets and reference values build on the overall understanding that the intervention, through causal pathways assumed by the intervention logic, ultimately contributes to broader policy goals. The indicators – be they input, output or outcome-oriented – are identified to support the evaluation of the respective interventions. Therefore, at least theoretically, through clear identification of evidence required for the assessment of contribution claims, both CA and CT could be made feasible.

From this starting point, the table below discusses the four dimensions listed above for both CA and CT approaches.

Table 1: Applicability of CA and CT to EU M&E

Dimension	Contribution analysis	Contribution tracing
Clarity of information generated for policy purposes	<p>Results of the analysis provide evidence of policy contribution to impacts. Current analyses are rather uncertain on this aspect. Therefore, CA is likely to improve clarity.</p> <p>However, the assessment is only qualitative and does not indicate magnitude.</p>	<p>Results of the analysis provide evidence of policy contribution to impacts. Current analyses are rather uncertain on this aspect; it would therefore improve clarity. The quantitative elements of the approach lend more credibility to outputs than CA, which would even improve this aspect.</p> <p>However, mathematical formulae and probabilities detract from the ease of interpretation.</p>
Data requirements and expertise needed for implementation	In terms of data requirements, depending on the specific intervention, it could be similar to the current framework, but could also be quite demanding for very complex interventions. The expertise needed would, after a period of adaptation, not differ widely from currently available/required knowledge.	In general similar to CA. However, the correct application of mathematical formulae would – especially at the beginning – require input from specialists and could be difficult to roll out across all programmes.
Plausibility of use for regular monitoring as opposed to one-off evaluation exercises	Given the considerable time required for evidence of the policy to materialise, only short-term results could form part of regular monitoring exercises. Data on longer-term outcomes and impacts would come with a considerable lag and would likely be only produced every couple of years.	Similar to CA.
Potential for simplification	The operationalisation of ‘evidence’ required to assess contribution claims could result in the use of many of the same as under the current framework.	Similar to CA, but due to its more demanding nature and need to quantify probabilities, could be more difficult to conceive as a regular monitoring exercise.

Source: Own elaboration

3.6. Towards the implementation of a TBE approach to social tracking

A comprehensive evaluation using a theory-based approach is a demanding exercise, likely to require far more effort and resources than can be routinely done by the BUDG Committee of the EP. This is especially so where there may be a premium on obtaining quick answers to justify policy choices or demand for change in how the policy is structured and implemented. Nevertheless, the principles behind TBE can provide a more informative picture of the performance of EU expenditure than relying exclusively or predominantly on a limited set of indicators. This section suggests some orientations for such an approach which, if endorsed in response to this study, can be fleshed out in more detail in the future.

As has been stressed repeatedly in this study, the impact of many – arguably, all – EU social spending interventions will not only take time, but also be determined by the effects of complementary policies undertaken by other agencies and levels of government. However, looking only at the connections between inputs and easily measurable outputs may well miss a key part of the story, namely whether the budgetary expenditure being tracked is heading in the right direction in the social outcomes it affects.

A plausible means of tracking can build on distinctions between outcomes to be expected at different stages of a process of change; these might be early or immediate outcomes, those occurring later in a process, or longer-term outcomes that then pave the way towards impacts. Making these distinctions will be facilitated by having a coherent theory of change/logic of intervention setting out both milestones (be it for inputs or outputs) and at least some sense of the underlying processes through which the desired change should arise.

Mixed methods are worth considering where the policy context is complex and there are multiple causes. Indicators may provide some answers, but it is more likely that only qualitative judgements will be viable. These might be merely binary – for example, ‘is there progress or not?’ – but will provide more nuanced and ambitious assessments if an ordinal scale is used. The latter could be, say, a five-point scale from ‘no progress or deterioration’ to ‘exceeding expectations’. For a set of policy priorities as wide-ranging as those in the EPSR, separate theories of change will unavoidably be needed and TBM cannot generalise across a policy spectrum.

Some methods can facilitate tracking. For example, known problems with earlier programmes, drawing on previous evaluations, can to some extent also be used to set benchmarks against which performance can be appraised. It may be useful to break down outcomes into those expected to be immediate, intermediate and ultimate; immediate is directly caused by outputs; intermediate only appears once immediate has been realised and can shape behaviour, for example. In the terminology of Figure 1, above, monitoring using a theory-based approach is, therefore, somewhere beyond outputs (i.e. in the outcome steps, but short of long-run impacts).

Then there is the matter of who is responsible for data collection and validation. Relevant data should have a clear source, and be of sufficient quality and reliability to track performance. Having a baseline is important, because it enables progress to be assessed. The use of targets, pathways and milestones will enable the EP to track what social expenditure is achieving (or not...).

Insights are also likely to be gleaned from examining the role of those implementing a policy based on EU expenditure. Quality of implementation can make a difference, but it will also be instructive to explore whether agents delivering a policy act as theory or assumptions imply. If they do not, there may be something to be learned about whether the original theory was flawed, whether the implementation is at fault and can be improved or, instead, whether an unanticipated deviation of the policy has occurred.

4. MAINSTREAMING SOCIAL OBJECTIVES IN EU SPENDING

From a broad perspective, mainstreaming is about the way policies, programmes and actions are formulated, designed, implemented and monitored, so as to take account of overarching societal goals and priorities. It may be supported by a number of tools and measures, at various stages of the programming cycle (e.g. integration into work programmes, screening, risk and impact assessments, management plans, capacity development and technical assistance). In the context of the EU budget, mainstreaming is defined as the inclusion of a specific priority in the design, preparation, implementation and evaluation phases of all budgetary programmes, in order to maximise their contribution to policy objectives, promote synergies and improve spending coherence. It may involve defining a specific target for the amount to be spent and monitoring progress towards achieving it (Levarlet et al., [2022](#)).

Mainstreaming common EU objectives in EU spending has only recently become a priority for the Commission to pursue, hand-in-hand with the shift towards performance-based budgeting. As explained above in Section 1.2, the Commission is currently tracking expenditure linked to the climate, biodiversity, clean air and migration. More recently, the Commission has provided qualitative information on gender equality and the common objectives in the SDGs. Yet a proper strategy for the mainstreaming of social objectives is lacking. This section looks first at the existing mainstreaming practices, then shifts the focus to social mainstreaming and investigation of the scope for mainstreaming the objectives of the EPSR in EU spending. The analysis builds on the recent experience of the EU Social Taxonomy proposed by the Social Taxonomy Subgroup of the EU Platform on Sustainable Finance in February 2022.

4.1. Existing EU mainstreaming objectives practiced in the EU budget

Although the EU currently has in place a climate and biodiversity and a gender mainstreaming methodology, there are notable differences in approach between the two. Climate and biodiversity mainstreaming relies on setting clear quantitative indicators that measure the green impact of EU spending. By contrast, the gender mainstreaming methodology proposes ‘procedural monitoring’ via key actors’ involvement in the implementation of EU spending. In what follows, the current climate mainstreaming and gender mainstreaming approaches are examined as illustrative cases.

4.1.1. Climate mainstreaming

In the 2014-2020 MFF, the Commission established a new approach, the so-called climate mainstreaming, to earmark financial resources for the fight against climate change. Programmes funded by the EU budget were obliged to allocate at least 20% of spending to climate change measures, targeted at achieving climate mitigation (reduction of greenhouse gas emissions) and climate adaptation (adapting to the effects of global warming). In December 2020, the Commission, the European Parliament and the Council added to the previous commitment by signing the Interinstitutional Agreement (IIA) (European Parliament, Council of the European Union and European Commission, [2020](#)), accompanying the agreement on the 2021-2027 MMF. The agreement consists of:

- i) an overall target of at least 30% for climate-relevant expenditure;
- ii) a ‘climate adjustment mechanism’, allowing for action to be taken if expenditure levels are likely to be insufficient to reach (programme-specific) climate spending targets;

- iii) the development of an effective climate tracking methodology to track the level of expenditure¹²; and
- iv) the application of the 'do no significant harm' (DNSH) principle to ensure that money spent under the budget does not prevent the EU from achieving its climate and environmental goals.

A Commission communication on the performance framework for the EU budget under the 2021-2027 MFF (European Commission, [2021d](#)) explains how climate mainstreaming operates as part of spending within the EU budget and the proposed methodology. Key improvements as compared to the 2014-2020 MFF (European Commission, [2022b](#)) are:

- i) a shift towards 'effect-based' expenditure tracking, based on a list of specific activities ('activity-based' tracking), compared to the more 'intent-based' approach in the 2014-2020 MFF;
- ii) improved consistency in marking expenditure across different funding programmes by reducing the risk that similar projects will be categorised differently depending on the objectives of the programme in question;
- iii) greater granularity and transparency in expenditure marking by introducing climate tagging in the Commission's official accounting and reporting systems; and
- iv) improved capacity to steer expenditure towards meeting both sector-specific and overall targets.

Additional elements were introduced in the new programming period in order to support climate mainstreaming, especially related to the DNSH principle (as outlined in Article 9 of Regulation (EU) 2021/1060 by the European Parliament and the Council ([2021](#))). Climate and environmental proofing of investments in infrastructure is thus required for all EU funds, with the aim of preventing financing for interventions identified by Article 17 of the Green Taxonomy (see the box below) as detrimental to the environment. The latter means that they produce significant negative environmental effects on the six environmental dimensions identified in the Green Taxonomy: climate change mitigation, climate change adaptation, sustainable use and protection of water and marine resources, the circular economy, pollution prevention and control, and the protection and restoration of biodiversity).

As part of the 2021-2027 climate mainstreaming architecture, the 30% target has been segmented into specific targets for individual programmes, each of them specifying expected contributions (as a percentage of the entire budget allocated) or making a more general commitment to climate mainstreaming (wherever it was not possible to define an ex ante spending target compatible with the objectives or nature of those programmes). The RRF, the ERDF and the Cohesion Fund imposed binding climate-spending targets intended to ensure a 'greener' programming (compared to the 2014-2020 MFF). Moreover, the related work programmes have to spell out how climate financing is proactively integrated into programming activities to attain the desired spending contribution.

¹² The term 'tracking' refers to the general approach of measuring and reporting climate expenditure, while 'tagging' refers to the action of assigning a specific coefficient to a certain intervention (European Commission, [2022b](#)).

Box 2: The Green Taxonomy

The introduction of a clear green mainstreaming methodology to embed environmental impact in EU spending was facilitated by the introduction of, and agreement on, the EU Taxonomy. This is a classification system which establishes a list of economic activities, with technical screening criteria that determine the conditions under which an activity is considered as environmentally sustainable. Technical screening criteria are defined by means of delegated acts by the Commission, whose scope is to identify activities that contribute substantially to achieving the EU's climate objectives and to set out 'do no significant harm', science-based, clear, practicable and easy-to-apply criteria. As such, the taxonomy is a system of classification and standard-setting, based on comprehensive sets of data, information and resources that can still be amended. The EU Taxonomy sets out four overarching conditions that an economic activity must meet to qualify as *environmentally sustainable* (Article 3):

- i. a significant contribution to achieving one or more of the EU's six stated environmental objectives defined in Article 9;
- ii. no significant harm to such objectives (Article 17);
- iii. compliance with minimum safeguards (Article 18); and
- iv. compliance with technical screening criteria (Articles 10-15).

In June 2021 the Commission adopted the first delegated act on sustainable activities for climate change adaptation and mitigation objectives (European Commission [2021e](#); [2021f](#); [2021g](#)). The Technical Expert Group also provided a list of NACE macro-sector categories, identified as the most relevant to climate objectives. A communication on '*EU taxonomy, corporate sustainability reporting, sustainability preferences and fiduciary duties: Directing finance towards the European green deal*' accompanied the first delegated act, with the aim of delivering key messages on how the sustainable finance toolbox facilitates access to finance for the transition (European Commission, [2021h](#)). A delegated act supplementing Article 8 of the Taxonomy Regulation was adopted in July 2021 by the Commission for scrutiny by the co-legislators. The act specifies the content, methodology and presentation of information to be disclosed by financial and non-financial undertakings concerning the proportion of environmentally sustainable economic activities in their business, investments or lending activities (European Commission, [2021i](#)). Finally, in February 2022, the Commission approved in principle a Complementary Climate Delegated Act (European Commission, [2022c](#), [2022d](#), [2022e](#), [2022f](#)), including, under strict conditions, specific nuclear and gas energy activities in the list of economic activities covered by the EU Taxonomy.

4.1.2. Gender mainstreaming

Gender mainstreaming is defined as 'the (re)organisation, improvement, development and evaluation of policy processes, so that a gender equality perspective is incorporated at all levels and at all stages, by the actors normally involved in policy-making' (Council of Europe, [1998](#): 13). More specifically, gender mainstreaming not only aims to avoid the creation or reinforcement of inequalities, but also implies analysing the existing situation, taking into account the different needs and interests of women and men, with the purpose of identifying inequalities. The aim is to underpin the development of policies which redress and undo the mechanisms that caused such inequalities. This requires both integrating a gender perspective into the content of the different policies and addressing the issue of the representation of women and men in the given policy area throughout.

At EU level, the Commission, the European Parliament, and the Council of the EU are in charge of implementing gender mainstreaming, whereas at national level it is up to the governments of Member States. On 5 March 2020, the Commission presented its Gender Equality Strategy 2020-2025 (European Commission, [2020b](#)). The strategy establishes policy objectives and actions to address gender equality challenges in Europe and 'seeks to include the gender perspective across all policy areas, at all levels and in all stages of policy-making'. In addition to concrete objectives and targets in the strategy, gender

mainstreaming requires a clear action plan. Such a plan should take into account the context, satisfy the necessary conditions, cover all the relevant dimensions, envisage the use of concrete methods and tools, set out the responsibilities, and make sure that the necessary competences exist to achieve the anticipated results within a planned timeframe.

A gender impact assessment is the first step towards avoiding policies that fail to consider a gender perspective. These assessments have to analyse the impact of a new regulation, policy or programme on the advancement of gender equality and anticipate implications for the lives of women and men. Moreover, by collecting information on the representation of both sexes as users/beneficiaries, the policy measures can be better targeted and their effects on different groups better estimated. Unbalanced participation of women and men in the planning and decision-making processes for policy actions may affect outcomes that affect both women and men. In order to integrate the gender perspective in the policy process, a multi-stage cycle is used.

The first step is the definition of specific policy needs that must be addressed by the public intervention in precise policy fields. In this stage, an assessment seeks to establish to what extent the policy is relevant for a defined objective and to understand what interventions are needed. To make valuable analyses and impact assessments, data and information have to be properly collected: they would contribute, along with the support of consulting stakeholders, to a strong understanding of the overall context, strengthening policy design and enhancing the learning process in the subject area. Once steps and milestones are defined and tasks and responsibilities are assigned, the plan is formalised and communicated. Then, when a first analysis is produced, the findings would help to define specific policy goals.

Second, once the policy objectives have been identified, precise interventions must be planned to achieve the goals set in the policy programme. During this stage, it is important to assess how budget allocations affect the stated policy objectives, as well as ensuring accountability and transparency in the way public funds are spent. When planning, it is vital to establish indicators that can contribute to the monitoring of the objectives, by measuring and comparing the policy or programme impact on the targeted population over the implementation period. Another important aspect is the inclusion of requirements in the evaluation criteria of proposals and tenders; they should be checked during the project implementation phase.

Third, the implementation stage must be closely followed, ensuring that all the involved participants are aware of the relevant objectives and plans, or, if this is not the case, that there is an obligation to set up briefings, capacity-building meetings and appropriate training. Care should also be taken to increase the traction among researchers, proposal evaluators, monitoring and evaluation experts, scientific officers, programme committee members, etc. Additionally, support measures during the implementation phase would produce further guidance. Ensuring that accountability mechanisms are in place to guarantee proper follow-up of implementation and progress is crucial. For this reason, it is important to envisage regular reporting and sharing of results.

4.2. Mainstreaming the EPSR objectives in the EU budget

These green and gender mainstreaming methodologies provide contrasting models for how to design a possible methodology for mainstreaming the EPSR's principles in EU spending. If a social mainstreaming methodology sought to follow the climate objective mainstreaming approach, it would have to measure the embedded social externalities/returns on public investment. By contrast, if an approach more similar to gender mainstreaming were considered better, a more procedural understanding of social mainstreaming would be required. What follows reviews the recent debate on an EU Social Taxonomy and reflects on possible insights and implications for the development of a social mainstreaming methodology in the EU budget.

4.2.1. Debate on the EU Social Taxonomy

In July 2021, the Commission charged the Social Taxonomy Subgroup of the EU Platform on Sustainable Finance with exploring an extension of the taxonomy to social objectives. The proposed EU Social Taxonomy is a tool to direct capital flows to activities that sufficiently respect human rights, and to investments that improve living and working conditions, especially for the disadvantaged. To build an EU Social Taxonomy, the Social Taxonomy Subgroup report takes the Green Taxonomy as a model by first defining the social objectives, then substantial contributions to achieving those objectives, and finally criteria that apply the DNSH principle.

The Social Taxonomy Subgroup also explicitly highlights important differences from the Green Taxonomy:

- i) the existence of inherent and additional social benefits in economic activities;
- ii) the fact that social impacts are not only derived from enlarging the production of certain goods and services, but also from respect for social (notably, labour) rights, which refers more to the economic entity than to the activity;
- iii) the fact that a social taxonomy cannot be based on scientifically validated research results;
- iv) the fact that it is more difficult to develop quantifiable criteria for a social taxonomy than for an environmental taxonomy.

With respect to the third and fourth points, a key strength of the Green Taxonomy is its capacity to identify environmentally sustainable activities, for instance based on quantifiable returns in terms of CO₂ emissions. The same, however, is not true for social returns, not even directly for social spending, let alone measuring the social returns of other types of activities. Very few studies have systematically analysed the social returns of public social spending, and findings are somewhat contradictory. Hemerijck et al. (2016), for instance, focus on the quantitative macro and micro impact of Early Childhood Education and Care (ECEC) and active labour market policies on employment and poverty. With respect to the former, the authors fail to find a correlation between ECEC spending and increased employment rates.

Bakker and Van Vliet (2019), who focus on social investment and the impact on employment, confirm that no statistically significant coefficient estimates for early childhood policies are obtained. Similarly, for efforts on education, they generally fail to obtain statistically significant effects. The lack of scientific evidence of the social returns of public social spending is linked – as also observed in Section 2 – to the complexity of measuring social externalities and isolating causality, which would require a context-specific understanding of the policy implementation that does not allow for generalisation. This methodological consideration explains, in large part, the difficulty of arriving at a consensus on these processes.

With these caveats in mind, the initial draft report of the EU Social Taxonomy Subgroup identified two dimensions for classifying activities based on their social sustainability¹³:

- A vertical dimension, which aims to promote adequate living standards and includes improving accessibility to products and services for basic human needs (e.g. housing, healthcare and education) and basic economic infrastructure (e.g. transport, telecommunication and clean electricity);
- A horizontal dimension, which aims to promote positive impacts and avoid/address negative impacts on affected stakeholder groups by ensuring decent work, promoting consumer interests and enabling inclusive and sustainable communities.

The vertical dimension focuses on the types of activities that could be considered 'social' based on the substantial contribution of such activities to the fulfilment of certain social objectives. 'Substantial contribution' here is operationalised based on the availability, accessibility, acceptability, and quality (AAAQ) approach. For the horizontal objectives, the focus shifts from the social characteristics of the economic activity (i.e. the products or goods being produced) to the (non-inherent) impact on people that is associated with the economic activity. The economic activity should not significantly harm the EU social objectives. The tables below present two illustrative examples of the application of the Social Taxonomy for the identification of eligible activities that contribute to the social objective 'adequate living standards and well-being for end-users', and to the specific sub-objectives 'access to healthcare' and 'access to housing'.

Table 2: Application of Social Taxonomy to adequate living standards and wellbeing for users (objective), notably in access to healthcare (sub-objective)

	Rationale for selecting sectors		
Selection of sectors	Sectors that provide products and services for basic human needs, according to the Universal Declaration of Human Rights accepted by the General Assembly of the UN on 10 December 1948 (Articles 11 and 25). Sectors related to the production of goods and services for basic human needs: food, water, housing, healthcare and education (European Social Pillar).		
Type of substantial contribution	Enhancing positive impact inherent in economic activity (examples)		
Substantial contribution	<p><u>Improve availability</u> Healthcare is available in sufficient quantity to cover people in need.</p> <p><u>Improve accessibility</u> Medication is affordable and economically sustainable for the countries in which it is distributed. There is an access strategy for new drugs, which is developed when it is clinically tested. There is monitoring of the percentage of the patient group that needs the medication, and that percentage is reached. The number of countries in which the drug is needed and filled for registration.</p>		
	Adequate living standards and wellbeing for end-users	Decent work	Inclusive and sustainable societies and communities
DNSH	Guarantee quality. Meeting internationally accepted standards of quality.	ILO core labour standards must be met. There must be adequate occupational safety and health (OSH). The minimum wage must be paid to workers.	Effluent must not pollute drinking water.

Source: Platform on Sustainable Finance (2022), p. 58.

¹³ These were then conflated in a single structure in the final report presented in February 2022.

Table 3: Application of Social Taxonomy to adequate living standards and wellbeing for users (objective), notably in access to housing (sub-objective)

	Rationale for selecting sectors		
Selection of sectors	Sectors that contribute to a reduction in the number of people without adequate, safe and affordable housing. Building and managing apartments and houses; NACE Code 41.20. Construction of residential and non-residential buildings.		
Type of substantial contribution	Enhancing positive impacts inherent in economic activity (examples)		
Substantial contribution	<p><u>Improving the availability/accessibility of housing units</u> Percentage of flats available for low income and disadvantaged people. Flats should have at least an x% cheaper rent than the average in the region, and access should be granted by means of a housing permit only. Percentage of m² assigned through a public waiting list.</p> <p><u>Improving accessibility</u> Build open spaces designed for social activities; foster access for vulnerable groups; ensure the presence of additional public services; ensure the existence of a social-management project.</p>		
	Adequate living standards and wellbeing for end-users	Decent work	Inclusive and sustainable societies and communities
DNSH	<p><u>Guarantee acceptability</u> Housing units/flats must follow the environmental taxonomy for the protection of biodiversity and ecosystems.</p> <p><u>Guarantee quality</u> Housing units/apartments must respect DNSH criteria set for the sector under the environmental taxonomy. There must be security of tenure for vulnerable groups. On process-related criteria, human rights and labour rights must be respected when building the units/apartments.</p>	<p>ILO core labour standards must be met. There must be adequate OSH. The minimum wage must be paid to workers.</p>	<p>No involuntary resettlement, and no pollution of drinking water when building apartments.</p>

Source: Platform on Sustainable Finance (2022), pp. 56-57.

Against this backdrop, to what extent can the Social Taxonomy be used for identifying socially sustainable investments and thus mainstreaming the EU social objectives in the EU budget? A first observation is that valuable insights can be taken from the vertical dimension of the Social Taxonomy. It identifies the sectors that are traditionally the responsibility of the state, such as the abovementioned housing, education and healthcare sectors, all of which obtain some financing from EU spending. Most importantly, the vertical dimension defines criteria for assessing the contribution of social expenditure financed by the EU budget to achieving social objectives. In the same vein, the horizontal DNSH principles provide a basis for guaranteeing a coherent framework for respect for adequate living standards and wellbeing for end-users, and decent work for people working in social services. In what follows, the possible application of the vertical and horizontal dimension to EU social spending is elaborated.

4.2.2. Application of the AAAQ criterion to EU social spending

In a recent study on the role of the RRF in strengthening childcare services, Corti et al. (2022) apply the AAAQ criterion to study the state of play in five countries, pre-COVID-19, and investigate how the RRF contributes not only to the construction of new places for children but also to the setting-up of an adequate, affordable, available and quality service. The key conclusion is that – despite the positive impact of EU funding on increasing, at the aggregate level, the availability of places – none of the countries under investigation introduced any change to increase the affordability, adequacy and quality of the service. The implication is that there is a risk of further aggravating rather than reducing inequalities in access to services, especially for low-income households. If an AAAQ criterion had been in place, the national plans on childcare could well have been revisited.

Applying such a criterion, however, cannot be done *ex ante*; instead, it requires a different step and thus a procedural understanding of mainstreaming. In this respect, the current gender mainstreaming approach offers a valuable example. To make sure EU social spending in social investment services is compliant with the AAAQ criteria, an in-depth understanding of the specific needs and problems of a certain policy would be required. This means that data and information have to be properly collected with the support of consulting stakeholders. On this basis, milestones would be defined, tasks and responsibilities assigned, and goals defined.

These will have to be regularly monitored to guarantee proper mainstreaming of the actions identified to improve the service's AAAQ. Regular monitoring is essential for social mainstreaming to ensure that all the involved participants are aware of the relevant objectives and plans or, if this is not the case, that there is an obligation to set up briefings, capacity building meetings and appropriate training. To sum up, social mainstreaming in EU spending could be applied to EU social spending by introducing an AAAQ criterion to trace the contribution of social spending programmes to the achievement of social objectives. As stressed above, such a criterion would work particularly well for social investment services, although not for traditional social protection cash benefits, none of which are financed through EU spending.

However, the application of the horizontal AAAQ criterion to EU social spending would be a demanding exercise. First, the definition itself of what is available, affordable, accessible and quality service is subject to discretion and ultimately depends on each country's institutional setting. Second, the need for an agreement on a monitoring framework for AAAQ services would lead to a heated political debate. Childcare is a good example of a policy area where – despite the existence of an EU framework and the demand for indicators, countries remain reluctant to use common indicators. Third, the AAAQ principle would be applied only to EU social spending, which would be contrary to the spirit of budgetary mainstreaming focusing on the inclusion of cross-cutting policy goals in all phases of the policy cycle (design, preparation, implementation and evaluation) of the relevant programmes, including those not strictly social.

Against this background, the application of a horizontal social mainstreaming clause based on the AAAQ criterion cannot readily be recommended.

4.2.3. Application of the social DNSH principle to EU spending

As stressed above, the main purpose of the Social Taxonomy is to serve as a guide for investors to direct capital flows towards (mainly non-social) activities subject to a set of social standards. Economic activities that comply with labour standards, including working conditions, non-discrimination, social protection, etc. are classified as 'socially sustainable'.

The DNSH criteria in a social taxonomy ensure that when an activity makes a substantial contribution to one social objective, it does not harm other social objectives. In other words, an activity, financed by the EU budget, that makes a substantial contribution to the objective on decent work should not harm end-users, communities and societies. For example, *'an economic activity such as broadband expansion in underserved areas makes a substantial contribution to the objective on inclusive and sustainable communities and societies. However, broadband expansion should not harm the rights of workers building the broadband infrastructure and the living standards and well-being of consumers using the internet services'* (Platform on Sustainable Finance (2022), p. 44). The Social Taxonomy, as illustrated above, identifies three main objectives to be monitored:

- i) Decent work: the formulation of this objective acknowledges the helpful role of the decent-work agenda spearheaded by the International Labour Organization (ILO). The four pillars of the decent work agenda are: employment creation, social protection, rights at work and social dialogue;
- ii) Adequate living standards and wellbeing for end-users: this objective focuses on people in their role as end-users of certain products and services that either pose heightened health or safety risks or that have the potential to help people to meet basic human needs;
- iii) Inclusive and sustainable communities and societies: this objective will emphasise respecting and supporting human rights by paying attention to the impacts of activities on communities and wider society.

Decent work is an objective directly related to EPSR principles: applying a social mainstreaming horizontal clause to EU-financed activities would imply qualifying as socially sustainable those contributing to the following sub-objectives:

- i) Promoting decent work, by:
 - a. Strengthening social dialogue, promoting freedom of association, and promoting collective bargaining for setting wages and working conditions;
 - b. Ensuring pay levels for workers are set in a predictable and transparent way;
 - c. Ensuring that formal working relationships avoid precarious working conditions
 - d. Providing excellent health and safety for workers;
 - e. Running extensive programmes for skills and life-long learning, job transition, and employment generation;
 - f. Providing for social protection as appropriate in the national context;
 - g. Taking immediate and effective measures to end forced labour and exploitation of work;
- ii) Promoting equality and non-discrimination at work through:
 - a. Equal employment opportunities for women, including by helping to end segregation in the labour market through gender-specific professions;
- iii) Ensuring respect for the human rights and workers' rights of affected workers in the value chain by carrying out risk-based due diligence including on the sub-objective areas listed above.

Applying a horizontal social mainstreaming principle to EU spending would imply guaranteeing that the objectives and sub-objectives of decent working conditions are applied horizontally to all activities financed through EU spending, and are compliant with such principles. Contrary to the vertical AAAQ

criterion, the DNSH principle could also be applied ex ante to any kind of EU spending, including non-social expenditure. As it is formulated in the Social Taxonomy proposal prepared by the Subgroup, a social DNSH would however apply to a limited subset of principles of the ESPR, notably those in ESPR Chapter 2 – ‘Fair working conditions’. Another limit of the DNSH is that it would function as a ‘social proofing’ procedure more than a mainstreaming practice¹⁴. As stressed above, mainstreaming is about the inclusion of cross-cutting policy goals in the EU budget. The application of a social DNSH principle would not be about EU spending to pursue social objectives, but rather guaranteeing that EU spending is compliant with EU social standards. Even though the former would be extremely useful, it would not ensure full mainstreaming of the EPSR objectives in the EU budget.

To sum up, this section explored the possible application of new social mainstreaming procedures in the EU budget. Drawing on insights from the current debate on an EU Social Taxonomy, two potential implementations of a vertical social mainstreaming procedure by means of an AAAQ criterion to social spending and a social DNSH clause to all EU spending in general were investigated. The analysis examined the potential pros and cons of each alternative and concludes that while a social mainstreaming methodology based on the principles of the EPSR has considerable potential, its prospective application needs further thought.

¹⁴ We are grateful to Federico Barbuto for this valuable suggestion.

5. CONCLUSIONS AND RECOMMENDATIONS

The core challenge in tracking social expenditure from the EU budget is that, although there is an abundance of information, it does not easily provide answers to a deeper policy question, which can be summed up as ‘do EU-funded policies contribute to achieving the desired societal transformations’?

The notion of the ‘logic of intervention’ is, therefore, central to providing answers, because it enables political decision-makers to conceptualise how they envisage a policy should work and then to assess whether it fulfils these expectations.

The problem, though, is that the impact of policies will only rarely be immediate and, instead, will more typically arise with a considerable lag. In addition, different stakeholders are likely to have different expectations. Thus, what a local agency charged with implementing a policy wants may differ considerably from the broader objectives of the European Parliament in its capacity not just as a budgetary authority, but also as an initiator of policy. More specifically, what an implementing authority wants will often reflect local priorities differing markedly from the broader thematic objectives set out for EU programmes, notably those embodied in the EPSR.

Because of this, any assessment of EU expenditure should start from a clear definition of the logic of intervention. An area as broad as social policy covers many initiatives, further complicating the adoption of effective tracking methods. The EPSR is an agreed EU policy framework, but even among the EU institutions there may be differences in perspective on the expected impacts and the timeframe over which they should be achieved. Hence, **a first important conclusion of this study is that the EP should try to articulate its own logic of intervention and what tracking it requires.** The logic should not only spell out the anticipated impacts, but also draw attention to how it differs from other stakeholders’. Working backwards, it is important to make explicit the manner in which the policy inputs supported by EU expenditure are expected to lead to direct outputs and, in turn, have outcomes which contribute to the desired impacts.

EU policies are only a part, and often only a small one, of the collective public sector effort to achieve social change. They can make a difference and may help to deal with a particular policy problem, but disentangling the specific contribution may be unrealistic. Sometimes, too, change occurs for reasons only tangentially related to policy interventions: as stressed in nearly all the interviews conducted for this study, it is captured in the phrase ‘**correlation is not causality**’. For these reasons, **a cautionary second conclusion of this study is that tracking may not yield convincing answers**, because it proves to be unable to measure change quantitatively. It may be that the best that can be expected is a more qualitative signal of whether or not, perhaps with some sense of extent, it improves the outcomes conducive to desired impacts. **Expectations of the quality of information that can be obtained through social tracking should therefore remain realistic.**

A further broad conclusion concerns the trade-off between demands for more intensive monitoring and excessive obligations for the relevant administrative bodies. Data collection has resource costs and evidence from interviews for this study suggests that these became excessive during the 2014-20 programming period. The **dilemma is not just whether monitoring obligations become too demanding, but also whether additional indicators bring added value.** The arguments for ‘more’ may be harder to justify for indicators available only with a considerable lag, or which do not evolve sufficiently rapidly to detect change soon enough to justify the alteration of policies. Yet there is **also a concern that reliance on a small number of easily collectable indicators reveals too little about policy effectiveness.** Moreover, the political difficulty in agreeing what these common indicators should be risks a ‘lowest common denominator’ result that satisfies no one. However, alternative approaches requiring the collection of data on programme implementation, with the involvement of various stakeholders and qualitative expert reviews, might be unduly burdensome and costly. These are awkward matters.

5.1. Existing monitoring and evaluation framework

The current M&E framework linked to social expenditure is already relatively well developed.

Significant changes were introduced during the 2014-2020 programming period, notably requiring more performance orientation and making the use of ex post evaluations mandatory across the policy cycle. An attempt to rationalise and consolidate this framework, based on feedback, has been undertaken for the current programming period. This includes, for instance, a reduced number of indicators to be reported on and the use of informed estimates (sampling) to ease the administrative burden of reporting.

However, it is not without its shortcomings. Despite these improvements, a number of challenges remain. Some of the common indicators are overly complex and the related data can be hard to collect. Data completeness is also a general issue, and data accuracy is difficult to check for Member States or the Commission. This is especially problematic when sensitive data (e.g. on programme participants' backgrounds) is concerned and data privacy has to be respected. Moreover, double-counting of certain outcome or result indicators seems to be a widespread practice across interventions.

Data collection is often perceived as an administrative burden by the authority concerned and not as an exercise to improve the quality of the policies. Indeed, some administrations are risk averse and therefore find it easier to account just for how much was spent, rather than whether the spending helps to meet an outcome target, let alone achieves the desired impact. What emerges from our analysis is that proper data collection is vital and should not be neglected, as it represents the foundation of an effective tracking mechanism. There is also a tension between having a performance tracking system and corresponding administrative burden, and the complaints of the Member States, often echoed in Parliament, that EU funding is too demanding. In addition, various interviewees questioned the utility itself of some indicators which – due to time lags in collecting them – might provide information that is already outdated.

The assessment of results and impacts is primarily undertaken at the ex post evaluation stage, although mid-term evaluations can provide valuable insights. Indicators in the current M&E framework mainly track inputs, outputs and more immediate results, primarily because of the difficulties and time lags highlighted above. This makes the assessment of policy impacts through monitoring alone very limited. Impact assessment is, instead, undertaken in ex post evaluations that aim to capture how the initial expectations were translated into results, and to what extent they have contributed to some observed impacts. In addition to theory-based evaluations, counterfactual impact evaluations – attempting to compare actual outcomes with what would have happened had the policy not been adopted – have been gaining increasing attention in recent years, especially for the ESF.

Care must be taken to ensure the clear and reliable identification of impacts generated through EU social expenditure. Analysis of causal mechanisms between specific instruments and observed impacts is particularly limited and, even where done, can be contested. This reduces the potential for evidence-based policymaking to introduce tailored improvements to future policy interventions. This limitation prompts questions about what tracking is intended to achieve and how realistic it might be to use evidence to justify a shift in the emphasis of policy.

The contribution of EU financing to social objectives cannot be fully disentangled from national-level funding programmes. Although the contribution of the Structural Funds and the RRF might be considerable in some Member States, overall they tend to be minor compared to national funding, especially in the social domains in which national and sub-national authorities are most prominent. Assessing the extent to which these different funding streams have led to the observed results and

impacts is highly problematic and risks distortion of policy messages. Even so, **ascertaining the added value of EU interventions should be a priority.**

Some of the **innovations introduced for monitoring the RRF can offer insights for the EP in social tracking**, notably relating to M&Ts. Milestones are defined in Article 2.4 of the RRF Regulation (EU 2021/241) as qualitative progress toward achievement of a reform or investment, while targets are quantitative. The introduction of this ‘financing not linked to costs’ approach should be welcomed as a positive step towards a new performance budgeting: a shift away – in the EU budget – from managing (i.e. ‘how much have we spent?’), to the achievement of policy objectives (i.e. ‘what have we accomplished with our money?’). The newly adopted paradigm is useful for policymakers since it provides them with valuable information that enables them to reassess policy priorities and target the kinds of programmes able to improve results and impacts.

Despite the positive nature of the RRF shift towards a new performance-based approach in EU budgeting, **some limits can be highlighted.** The first regards the rigidity of the current system of M&Ts. M&Ts are set *ex ante*. Due to inherent uncertainty and rapidly changing conditions in policymaking, this is no easy task. Modern investment practice often relies on smaller milestones and regular joint review to allow for revision, in line with an ‘experimentalist’ logic. This is not true of the RRF, which does not allow for a revision of the M&Ts.

Second, M&Ts rely mainly on national level indicators to track progress. Such national level indicators may miss the regional dimension and other contextual factors, such as the variety of institutional mechanisms, which may influence impact. Third, the plans were typically formulated at short notice and in a centralised manner. To speed up the processes, the M&Ts were in some cases only quickly discussed within the relevant line ministries before being agreed upon and negotiated with the Commission by the ministries of finance. Since the governing principle in the RRF implementation is ‘comply or don’t get paid’, Member States might be tempted to commit to easily attainable M&Ts, also including in the plans reforms and investments already planned to be sure to receive the Commission disbursement.

The selection of overly aggregated or ‘guaranteed’ milestones could be seen as a weakness in the RRF monitoring system. Indeed, even though a Member State is compliant with the indicators agreed with the Commission, and thus will successfully fulfil the M&Ts specified in the Council Implementing Decision (CID) and receive the Commission disbursement, this information may reveal little about the actual utility of the measures implemented vis-à-vis their objectives.

5.2. Data and indicator problems

For the great majority of indicators, data collection at EU level relies on that by national bodies. It follows that assurance of quality and reliability is not easy. **Different social spending programmes contributing to several objectives across diverse national and regional contexts makes EU-level aggregation of specific monitoring indicators difficult, and sometimes impossible.** It may even be that aggregation is the wrong approach, to the extent that diverse needs call for different interventions across Member States. The distinct nature of these programmes is also captured by the plethora of different programme-specific indicators that measure how they contribute to the achievement of the goal they were meant to address. While such specific indicators are important for measuring performance, they also reduce the potential for comparability.

Ensuring a common definition of indicators is critical for comparability. With 27 different national implementation systems and a very large number of Managing Authorities and Monitoring Committees in charge of designing and implementing OP-level data collection activities, a common

understanding of the type of information to be gathered is critical. A lack of such understanding calls the comparability of data into question. An additional difficulty, despite agreed definitions and conventions, is the lack of a rigorous and enforceable methodology for data collection, further hampering the collection of comparable information. Sometimes, the national definitions of certain indicators may even change over time. Besides making EU-level aggregation nearly impossible, low comparability constrains analytical efforts to assess the effectiveness of the respective programmes. To deal with these problems, **an analysis of the comparability of administrative data should be considered**. It would be a prerequisite for arriving at a common set of indicators and would facilitate reaching agreement by discarding options for which comparability is lacking.

Lags in the availability of result and impact variables are especially difficult in the domain of social policy. Therefore, programme effectiveness can only reasonably be assessed well after the specific programmes have ended. **The more focus there is on impact indicators that allow the success of a policy to be judged, the more difficult it is to link current spending and policy initiatives to results and impacts**. This is due to a variety of factors, making it difficult to differentiate between correlation, causality and incidental effects. An additional challenge for many social policies is that results tend to be intangible: the **construction of bridges or kilometres of rail is visible and quantifiable; advances in social inclusion may only be established qualitatively**.

EU social spending supports both social infrastructural investments (traditionally financed via the ERDF) and a variety of social programmes, such as food banks to support the most deprived, training for the long-term unemployed or counselling activities for young people not in employment, training or education. **A different mentality is needed when dealing with people as opposed to social infrastructure** (think about people in social exclusion). An important distinction should consequently be drawn between new buildings or equipment and social outcomes. Overall, too little attention is paid to the results of ESI Funds, and much more focus is placed on how the money is spent. Even output indicators are often derived from surveys, not reliable administrative data.

5.3. Mainstreaming social objectives in the EU budget

As a complementary avenue to social tracking, budgetary mainstreaming is about how policies, programmes and actions are formulated, designed, and implemented and monitored, taking account of overarching societal goals and priorities. Since mainstreaming common objectives (e.g. green and gender) in EU spending has only recently become a priority to pursue, this study has explored the scope for social mainstreaming to ensure that all EU-funded policies are compliant with the social standards defined in the EPSR.

Drawing on insights from the current debate on the EU Social Taxonomy, two different approaches to social mainstreaming are explored: one vertical and one horizontal. The first consists in guaranteeing that EU-funded social spending is compliant with common EU social objectives. The second is about making sure that non-social EU spending does not significantly harm EU social objectives.

With respect to the first approach, **an option would be to apply the availability, accessibility, acceptability and quality (AAAQ) principle to all social spending**. This approach would be applied in two steps. First, ex ante, social spending under the EU budget should be compliant with the four criteria and their approval should be conditional on meeting them. To this end, data and information must be properly collected, drawing on and consulting relevant stakeholders, and a monitoring system should be set up ex ante. On this basis, milestones would be defined, tasks and responsibilities assigned, and goals defined. They would have to be regularly monitored to guarantee proper mainstreaming of the actions identified to improve the service's AAAQ. **Regular monitoring is**

essential for social mainstreaming to ensure that all the involved participants are aware of the relevant objectives and plans or, if not feasible, that there is an obligation to set up briefings, capacity-building meetings and appropriate training.

Moving towards an AAAQ approach is not an easy task for at least three reasons:

1. The definition itself of what is available, affordable, accessible and quality service is subject to discretion and ultimately depends on each country's institutional setting;
2. The need for an agreement on a monitoring framework for AAAQ services would be the subject of a heated political debate;
3. The AAAQ principle would be applied solely to EU social spending, which would in part represent a limit to the very idea of budgetary mainstreaming, which is about the inclusion of cross-cutting policy goals in all phases of the policy cycle of the relevant programmes, including those not strictly social.

Moving to the horizontal dimension, this study explores the possible introduction of a social Do Not (socially) Significantly Harm procedure to be applied to all EU spending. Concretely, this means that any EU spending should be checked to avoid violating social objectives. Based on the Social Taxonomy proposal, three objectives to be monitored (which can be further expanded) were identified, reflecting the principles of EPSR Chapter 2 – 'Fair working conditions': i) promoting decent work; ii) promoting equality and non-discrimination at work; iii) ensuring respect for human rights and workers' rights. As in the case of the vertical mainstreaming, the limits of a social DNSH principle should be recognised:

1. As it is formulated in the Social Taxonomy proposal, a social DNSH principle would apply to a limited subset of principles of the EPSR;
2. It would function as a 'social proofing' procedure more than a mainstreaming practice, i.e. the application of a social DNSH principle would not be about EU spending to pursue social objectives but rather guaranteeing that EU spending is compliant with EU social standards.

As a general proposal, and taking into consideration the abovementioned pros and cons, there are sound reasons for making more extensive use of social mainstreaming practices in the EU budget. Doing so would contribute to embedding the principles of the EPSR in any EU spending, be it social or not. However, the way in which to do so remains contentious; the recommendation is therefore to explore in more detail how mainstreaming might be made both politically acceptable and possible to implement in a credible manner. In this respect, horizontal and vertical mainstreaming approaches should not be seen as alternative options. Because they have different objectives they can, instead, be seen as complementary. At the same time, one should not underestimate the degree of complexity which is implied in putting in place a new mainstreaming instrument, with the related administrative burden. Similarly, one should also consider the political priorities when deciding to introduce a new mechanism. Given that green and gender mainstreaming approaches are already in place, the EP should consider whether an additional social mainstreaming tool is politically desirable.

5.4. Recommendations

For the EP to be able to monitor whether the EU budget is being spent not just correctly (a task undertaken in tandem with the European Court of Auditors), **but also well**, it needs to have timely information. **Evaluations undertaken after a programme has been completed provide valuable information on how to design future programmes and pitfalls to avoid, but cannot be used to ascertain whether a programme in progress is, or is not, achieving its aims** and should be revised.

The move towards performance budgeting is a step in the right direction, but more is needed to shift the focus of tracking towards what policies achieve. It should be stressed in this context that while meeting formal legal obligations to achieve M&Ts as part of performance budgeting for RRF disbursements represents a significant innovation, it does not necessarily mean that the policy will achieve the impacts envisaged. Moreover, a key question for the budgetary authorities is **whether EU added value can be assured when EU initiatives are combined with national policies**.

Greater attention should also be paid to the implementation dimension of programmes, considering that even a well-conceived initiative could disappoint if poorly executed. Monitoring **processes as well as output and outcomes** should be part of assessing the potential impact.

Building on the TBE approach, EU social spending would be assessed not based on the annual/biannual achievement of M&Ts' outputs or on costs occurred, but rather **based on the qualitative assessment of the effectiveness and utility of the interventions**. This requires a combination of data collection on monetary inputs at the disaggregated level to account for the cross- and intra-regional distribution of resources, as well as a significant ex ante knowledge of what to expect from social expenditure. This includes understanding of projected socio-economic returns, awareness of the micro-foundations of social change (i.e. micro-outcomes in life course transitions), complementarities and sequencing of reforms trajectories, and understanding of the institutional mechanisms which have a bearing on national and subnational delivery of social policies.

Such an approach would yield a number of potential benefits. For national authorities, there would be some easing of the current administrative burden in reporting on output indicators, limiting problems in defining not only common indicators, but also outputs and results. In parallel, it would help to increase other stakeholders' involvement in programme monitoring and allow for flexible adjustments. For the European Commission it would deter unproductive box ticking in monitoring the implementation of programmes, but also pave the way for mid-term evaluations of the MFF/RRF, while also increasing transparency and flexibility.

The recommendations which follow distinguish between proposals to strengthen the system currently in place and overcome the limits identified in tracking social objectives, and proposals for a new experimental approach to social spending tracking. For the latter, **an overarching recommendation is to 'experiment' by selecting a specific area of social policy**. Selecting a specific area of social policy would make sense, considering that a one-size-fits-all approach would be ill advised, given the specificities of each programme. Conducting such a 'pilot study' is good practice before investing in a more comprehensive approach.

5.4.1. Strengthen the current M&E system to improve social tracking

Ensure coherence in the definitions of key concepts and indicators. Harmonising indicators across 27 Member States is a very challenging task, but one which would yield considerable benefits for the comparability of indicators collected. In the medium and long run, the objective should not be to ask for more indicators but rather to achieve greater coherence. In the short run, however, much could already be achieved by ensuring that definitions are strictly adhered to across EU legislation related to expenditure programmes. Despite the inevitable difficulties, further enhancement of consistency should be pursued.

The contribution of EU funds to social objectives should, in particular, be assessed on their added value vis-à-vis national funds, and as far as possible, on whether they act on problems neglected by other interventions. As the specific contribution of EU financing in national contexts could be impossible to assess in comparison to much more substantial national (and sub-national)

programmes, examining how they are complementary could provide valuable insights. A more ambitious approach could be to harmonise key elements of national and EU M&E frameworks to make these dimensions easier to assess. This is likely to increase the complementarity of the two funding streams as well, but clearly involves coordination and resource demands.

Further encourage the use of administrative data by managing authorities for monitoring and evaluation purposes. This would reduce the administrative burden on intermediaries and beneficiaries of EU funds by making use of existing data sources that often provide far more detail than the current monitoring approach. This would, in turn, allow for more detailed analyses to be produced and for an improved evidence-base for future policymakers. Leveraging this information is particularly relevant when monitoring the results of social interventions. While respecting the GDPR restrictions, as much access as possible to microdata should be enabled.

To foster greater consistency in data and to improve exchange of experience in social tracking, the **development of a framework for learning evaluation should be given priority.** Among the steps that could be taken to achieve this would be stimulating the creation of a policy community bringing together experts and practitioners, and working towards a common language and learning culture in the science-for-policy community. Two benefits can directly result from this virtuous process: peer-review and comparability, which should not be confused with attempts to harmonise principles. Capacity building in how to undertake social tracking and generating evidence on causal mechanisms would be worthwhile, especially in appraising what works (or does not). In addition, boosting access for experts to microdata already collected for administrative purposes should be considered. In addition to this, the acknowledgement of a functioning tracking system would already raise awareness among national authorities and actors involved in the data processing, as well as fund beneficiaries.

A partnership agreement should be drawn up with regional authorities to guarantee that regional and local actors are at the core of the monitoring and evaluation process. Such an agreement should be prepared in dialogue with the Commission and be based on procedures that are transparent for the public and in accordance with the institutional and legal framework of the Member States. Involvement of regional authorities – as well as social partners – in the identification of territorial specific indicators to track social spending and evaluate performance could improve the capacity for assessing the actual contribution of EU spending. The EP has long argued in favour of this partnership approach in EU funding, not only for the structural funds but also for the RRF, and it is important that it continues to push for the involvement of social partners, and local and regional authorities.

5.4.2. Launch a pilot project on TBE for selected social policies

Encourage the establishment of a theory-based tracking system that monitors the causal mechanisms leading to results and impacts. A plausible means of tracking can build on distinctions between outcomes to be expected at different stages of a process of change. These could be early or immediate outcomes, those occurring later in a process, and longer-term outcomes that then pave the way towards impacts. Qualitative assessments that indicate whether the direction of change is aligned with expectations – based on a rigorous and common methodology – could **allow progress towards impacts consistent with the logic of intervention to be measured, provided they are defined with care.** Milestones and progress towards targets are already the means by which decisions on disbursements of funds are determined, but the aim here would be to conduct a parallel exercise based on information available.

It would not be tied to funding, but would rather be about understanding at a more abstract level whether programmes were achieving their political objectives. In other words, **the goal would be to explore how well a policy is working, rather than tying monitoring to formal legal obligations.**

Concretely, the selected programmes would be assessed not based on common indicators or the annual/biannual achievement of M&Ts, but rather on the qualitative assessment of the effectiveness and utility of the interventions. It is important to stress that following this route would not be in conflict with legal obligations related to the monitoring of performance, but would complement these obligations to inform current and future policy strategies. In terms of **potential benefits** at both national and EU level, a TBE system would:

- National authorities:
 - Lessen the problem of defining common indicators and using them to assess outputs or results by concentrating on the effectiveness of interventions from the perspective of making progress on delivering the objectives of the EPSR;
 - Provide insights into determinants of programme performance, enabling the logic of intervention to be appraised; and
 - Increase the range of stakeholders engaged in monitoring programmes and allow for flexible adjustments based on evidence.
- European Commission:
 - Avoid what risks being a 'box-ticking' exercise in monitoring the implementation of programmes; and
 - Inform the mid-term evaluation of the RRF, while contributing to increased transparency and flexibility.
- European Parliament:
 - Play a proactive role in the monitoring of the RRF plans by drawing the Commission's attention to potential deviations from the initially stated social objectives and – in principle – suggesting corrections.

Our proposal is to **launch such a pilot project in parallel with existing RRF procedures and in view of the mid-term revision** of the Facility. The launch of this pilot project would allow for an initial experiment to be conducted in order to gain a more dynamic understanding of performance outcomes and a more flexible understanding of performance objectives. It would also be of interest to launch a separate pilot project covering the full policy cycle of intervention, from design through to implementation.

REFERENCES

- Acemoglu, D. (2010), *Theory, General Equilibrium, and Political Economy in Development Economics*, The Journal of Economic Perspectives, vol. 24(3), pp. 17–32, Available [here](#).
- Adato, M. (2008), *Combining survey and ethnographic methods to improve evaluation of conditional cash transfer programs*, International Journal of Multiple Research Approaches, vol. 2(2), pp. 222–236, Available [here](#).
- Arrow, K. J. (2006), *Freedom and Social Choice: Notes in the Margin*, Utilitas, Cambridge University Press, vol. 18(1), pp. 52–60, March, Available [here](#).
- Atkinson, T., Cantillon, B., Marlier, E. & Nolan, B. (2002), *Social Indicators: The EU and Social Inclusion*, OUP Catalogue, Oxford University Press, n. 9780199253494, Available [here](#).
- Bakker, V. & Van Vliet, O. (2019), *Social Investment, Employment Outcomes and Policy and Institutional Complementarities: A Comparative Analysis across 26 OECD countries*, Department of Economics and Institute of Public Administration, Leiden University, MPRA Paper No. 96140, University Library of Munich, Germany. Available [here](#).
- Bamanyaki, P.A. & Holvoet, N. (2016), *Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives*, Evaluation, vol. 22(1), pp. 72–90, Available [here](#).
- Befani, B. & Mayne, J. (2014), *Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation*, IDS Bulletin, vol. 45(6), pp. 17–36, Available [here](#).
- Bokhorst D., Corti, F. (2023), *Governing Europe’s Recovery and Resilience Facility: Between Discipline and Discretion*, Government and Opposition (forthcoming).
- Bouckaert, G., Peters, B. & Verhoest, K. (2010), *The Coordination of Public Sector Organizations: Shifting Patterns of Public Management*, Palgrave Macmillan UK, 326 pp., Available [here](#).
- Cejudo, G.M. & Michel, C.L. (2017), *Addressing fragmented government action: coordination, coherence, and integration*, Society of Policy Sciences, vol. 50(4), pp. 745–767, December, Available [here](#).
- Corti, F., Morabito, C., Ruiz, T. & Luongo, P. (2022), *The role of the Recovery and Resilience Facility in strengthening childcare policies*, Recovery Watch Policy Study, July, Available [here](#).
- Corti, F. & Núñez Ferrer, J. (2021), *Steering and monitoring the Recovery and Resilience Plans. Reading between the lines*, Recovery and Resilience Reflection Papers No. 2, CEPS, Available [here](#).
- Council of Europe (1998), *Gender Mainstreaming: Conceptual framework, methodology and presentation of good practices*, Strasbourg, May, Available [here](#).
- Darvas (2022), *The puzzle of European Union recovery plan assessments*, Bruegel post, Available [here](#).
- Devereux, S., Roelen, K., Béné, C., Chopra, D., Leavy, J. & McGregor, J.A. (2013), *Evaluating Outside the Box: An Alternative Framework for Analysing Social Protection Programmes*, IDS Working Papers, vol. 2013(431), pp. 1–26, Available [here](#).
- ESF Data Support Centre (2021), *ESF+ Common Indicators Toolbox Working Document*, Programming Period 2021–2027 Monitoring and Evaluation of European Cohesion Policy - European Social Fund Plus (ESF+) Shared Management Strand, Available [here](#).
- European Commission (2022a), *Mainstreaming in the EU budget*, European Commission website, Available [here](#).

- European Commission (2022b), *Climate Mainstreaming Architecture in the 2021- 2027 Multiannual Financial Framework*, Staff Working Document (2022) 225 final, Available [here](#).
- European Commission (2022c), *Commission Delegated Regulation (EU) amending Delegated Regulation (EU) 2021/2139*, C(2022) 631/3, Available [here](#).
- European Commission (2022d), *Annex 1 to the Commission Delegated Regulation (EU) amending Delegated Regulation (EU) 2021/2139*, C(2022) 631/3, Available [here](#).
- European Commission (2022e), *Annex 2 to the Commission Delegated Regulation (EU) amending Delegated Regulation (EU) 2021/2139*, C(2022) 631/3, Available [here](#).
- European Commission (2022f), *Annex 3 to the Commission Delegated Regulation (EU) amending Delegated Regulation (EU) 2021/2139*, C(2022) 631/3, Available [here](#).
- European Commission (2021a), *Better Regulation Toolbox*, November, Available [here](#).
- European Commission (2021b), *Commission Delegated Regulation (EU) 2021/2106 on supplementing Regulation (EU) 2021/241 of the European Parliament and of the Council establishing the Recovery and Resilience Facility*, Available [here](#).
- European Commission (2021c), *Commission Delegated Regulation (EU) 2021/2105 on supplementing Regulation (EU) 2021/241 of the European Parliament and of the Council establishing the Recovery and Resilience Facility*, Available [here](#).
- European Commission (2021d), *Communication from the Commission to the European Parliament and the Council on the performance framework for the EU budget under the 2021-2027 MFF*, COM(2021) 366 final, Available [here](#).
- European Commission (2021e), *Commission Delegated Regulation (EU) supplementing Regulation (EU) 2020/852 of the European Parliament and of the Council*, C(2021) 2800 final, Available [here](#).
- European Commission (2021f), *Annex 1 to the Commission Delegated Regulation (EU) supplementing Regulation (EU) 2020/852 of the European Parliament and of the Council*, C(2021) 2800 final, Available [here](#).
- European Commission (2021g), *Annex 2 to the Commission Delegated Regulation (EU) supplementing Regulation (EU) 2020/852 of the European Parliament and of the Council*, C(2021) 2800 final, Available [here](#).
- European Commission (2021h), *EU Taxonomy, Corporate Sustainability Reporting, Sustainability Preferences and Fiduciary Duties: Directing finance towards the European Green Deal*, COM(2021) 188 final, Available [here](#).
- European Commission (2021i), *Commission Delegated Regulation (EU) supplementing Regulation (EU) 2020/852 of the European Parliament and of the Council*, C(2021) 4978 final, Available [here](#).
- European Commission (2021j), *Guidance to Member States Recovery and Resilience Plans*, SWD(2021) 12 final, Brussels, Available [here](#).
- European Commission (2020a), *Evaluation of the ESF and YEI Support to Youth Employment*, Staff Working Document (2020) 216 final, Available [here](#).
- European Commission (2020b), *A Union of Equality: Gender Equality Strategy 2020-2025*, COM(2020) 152 final, Available [here](#).

- European Commission (2018a), *Programming period 2014-2020: Monitoring and Evaluation of European Cohesion Policy - European Social Fund Guidance document*, Directorate-General for Employment, Social Affairs and Inclusion, Available [here](#).
- European Commission (2018b), *Study on the monitoring and evaluation systems of the ESF: final report*, Directorate-General for Employment, Social Affairs and Inclusion, Available [here](#).
- European Commission (2018c), *Synthesis report of ESF 2017 annual implementation reports: final report*, Directorate-General for Employment, Social Affairs and Inclusion, Available [here](#).
- European Commission (2014), *Programming period 2014-2020: Monitoring and Evaluation of European Cohesion Policy - European Social Fund Guidance document*, Directorate-General for Employment, Social Affairs and Inclusion, Available [here](#).
- European Court of Auditors (2022), *Cohesion and NextGenerationEU: concord or clash?*, Journal n.1/2022, Available [here](#).
- European Court of Auditors (2021a), *Performance-based financing in Cohesion policy: worthy ambitions, but obstacles remained in the 2014-2020 period*, Special Report 24/2021, Available [here](#).
- European Court of Auditors (2021b), *ESF support to combat long-term unemployment: Measures need to be better targeted, tailored and monitored*, Special Report 25/2021, Available [here](#).
- European Court of Auditors (2020a), *Combating child poverty – Better targeting of Commission support required*, Special Report 20/2020, Available [here](#).
- European Court of Auditors (2020b), *Opinion No 6/2020 concerning the proposal for a regulation of the European Parliament and of the Council establishing a Recovery and Resilience Facility (COM(2020) 408)*, Available [here](#).
- European Court of Auditors (2019), *Delivering performance in Cohesion (Briefing paper)*, Review 08/2019, Available [here](#).
- European Court of Auditors (2018), *Selection and monitoring for ERDF and ESF projects in the 2014–2020 period are still mainly outputs-oriented*, Special Report 21/2018, Available [here](#).
- European Court of Auditors (2017), *The Commission's negotiation of 2014-2020 Partnership Agreements and programmes in Cohesion*, Special Report 2/2017, Available [here](#).
- European Parliament, Council of the European Union & European Commission (2021), *Interinstitutional Agreement on budgetary discipline, on cooperation in budgetary matters and on sound financial management, as well as on new own resources, including a roadmap towards the introduction of new own resources*, Available [here](#).
- European Parliament & Council of the European Union (2021a), *Regulation (EU) 2021/1057*, Available [here](#).
- European Parliament & Council of the European Union (2021b), *Regulation (EU) 2021/241*, Available [here](#).
- European Parliament & Council of the European Union (2013), *Regulation (EU) 1303/2013*, Available [here](#).
- Gilson, R., Sabel, C. and Scott R. (2009), *Contracting for innovation: vertical disintegration and interfirm collaboration*, Columbia Law Review, vol. 109(3), Available [here](#).

- Halpern, C., Jacquot, S. & Le Galès, P. (2008), *A Mainstreaming: Analysis of a Policy Instrument*, New Modes of Governance, Policy Brief n. 33, Available [here](#).
- Head, B.W. & Alford, J. (2013), *Wicked Problems: Implications for Public Policy and Management*, Administration & Society, vol. 47(6), pp. 711-739, Available [here](#).
- Hemerijck, A., Burgoon, B., Di Pietro, A. and Vydra, S. (2016), *Assessing Social Investment Synergies (ASIS)*, European Commission, Directorate-General for Employment, Social Affairs and Inclusion, Available [here](#).
- Hermans, L.M., Naber, A.C. & Ensenrink, B. (2012), *An approach to design long-term monitoring and evaluation frameworks in multi-actor systems - A case in water management*, Evaluation and Program Planning, vol. 35(2), pp. 427-438, Available [here](#).
- HM Treasury (2020a), *Magenta Book: Central Government guidance on evaluation*, United Kingdom Government, March, Available [here](#).
- HM Treasury (2020b), *Magenta Book Annex A: Analytical methods for use within an evaluation*, United Kingdom Government, March, Available [here](#).
- Istituto per la Ricerca Sociale (2016), *Mainstreaming Employment and Social Indicators into Macroeconomic Surveillance*, Study requested by the EMPL committee, European Parliament, Available [here](#).
- Krause, P. (2010), *M&E Systems and the Budget*, World Bank Publications - Reports 11077, The World Bank Group, Available [here](#).
- Laffan, B. & Schlosser, P. (2016), *Public finances in Europe: fortifying EU economic governance in the shadow of the crisis*, Journal of European Integration, vol. 38(3), pp. 237-249, Available [here](#).
- Lecerf, M. (2019), *European Social Fund Plus (ESF+) 2021-2027*, European Parliamentary Research Service Briefing, Available [here](#).
- Levarlet, F., Alessandrini, M., Schratzenstaller-Altzinger, M. & Franceschelli, N. (2022), *Climate mainstreaming in the EU budget: 2022 update*, Study requested by the BUDG committee, European Parliament, Available [here](#).
- Mause, K. (2019), *Citizens attitudes toward the EU budget: an overview*, in: Zamparini, L. & Villani-Lubelli, U. (ed.), *Features and Challenges of the EU Budget*, chapter 5, pp. 60-72, Edward Elgar Publishing, Available [here](#).
- Mayne, J. (2012), *Contribution analysis: Coming of age?*, Evaluation, vol. 18(3), pp. 270-280, Available [here](#).
- Mazey, S. (2000), *Introduction: Integrating gender - intellectual and 'real world' mainstreaming*, Journal of European Public Policy, vol. 7(3), pp. 333-345, Available [here](#).
- McConnell, J. (2019), *Adoption for adaptation: A theory-based approach for monitoring a complex policy initiative*, Evaluation and Program Planning, vol. 73, pp. 214-23, Available [here](#).
- Nunan, F., Campbell, A. & Foster, E. (2012), *Environmental mainstreaming: the organisational challenges of policy integration*, Public Administration and Development, vol. 32(3), pp. 262-277, Available [here](#).
- OECD (2008), *Performance Budgeting: A Users' Guide*, Policy Brief, Available [here](#).
- Platform on Sustainable Finance (2022), *Final Report on Social Taxonomy*, February, Available [here](#).

- Remme, M., Vassall, A., Lutz, B., Luna, J. & Watts, C. (2014), *Financing structural interventions: going beyond HIV-only value for money assessments*, AIDS, vol. 28(3), pp. 425–434, Available [here](#).
- Robinson, M. & Last, D. (2009), *A Basic Model of Performance-Based Budgeting*, IMF Technical Notes and Manuals 2009/001, International Monetary Fund, Available [here](#).
- Roelen, K. & Devereux, S. (2014), *Evaluating Outside the Box: Mixing Methods in Analysing Social Protection Programmes*, CDI Paper 6, Institute of Development Studies, Available [here](#).
- Samson, M., van Katwyk, S., Fröling, M., Ndoro, R., Meintjes, C., Buts, L. & Renaud, B. (2015), *Methods of measuring the impacts of social policy in political, economic and social dimensions*, UNRISD Working Paper No. 2015-4, United Nations Research Institute for Social Development, Geneva, Available [here](#).
- Sapała, M. (2018), *Performance budgeting: A means to improve EU spending*, European Parliamentary Research Service In-Depth Analysis, European Parliament, Available [here](#).
- Schick, A. (2014), *The metamorphoses of performance budgeting*, OECD Publishing, vol. 13(2), pp. 49-79, Available [here](#).
- Sen, A.K. (2009), *The Idea of Justice*, The Belknap Press of Harvard University Press, Cambridge, Massachusetts, 467 pp., Available [here](#).
- Smith, F. (2021), *The Welsh Government's use of policy tools for mainstreaming equalities*, Wales Centre for Public Policy, June, Available [here](#).
- Vanhercke, B., Verdun, A., Atanasova, A., Spasova, S. & Thomson, M. (2021), *From the European Semester to the Recovery and Resilience Facility. Some social actors are (not) resurfacing*, etui Working Paper 2021.13, Available [here](#).
- White, H. (2008), *Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation*, IDS Bulletin, vol. 39(1), pp. 98-109, Available [here](#).
- Yamin, F. (2005), *The European Union and future climate policy: Is mainstreaming adaptation a distraction or part of the solution?*, Climate Policy, vol. 5(3), pp. 349-361, Available [here](#).
- Zamparini, L. & Villani-Lubelli, U. (2019), *Introduction*, in: Zamparini, L. & Villani-Lubelli, U. (ed.), *Features and Challenges of the EU Budget*, chapter 1, pp. 1-10, Edward Elgar Publishing, Available [here](#).

LIST OF INTERVIEWS

1. BusinessEurope, 24/08/2022, Brussels (video conference, MTeams)
2. European Commission, DG-EMPL, 24/08/2022, Brussels (video conference, MTeams)
3. International Labour Organization, 25/08/2022, Brussels (video conference, MTeams)
4. European Committee of Regions, 30/08/2022, Brussels (video conference, MTeams)
5. European Commission, DG-ECFIN, 31/08/2022, Brussels (video conference, Zoom)
6. European Court of Auditors, 15/09/2022, Brussels (video conference, MTeams)
7. Head of service of the UAFSE – Galicia (Spain) ESF Operational Programme, 20/09/2022, Brussels (video conference, MTeams)
8. European Commission, DG-REGIO, 06/10/2022, Brussels (video conference, MTeams)
9. Lombardia region (Italy) – ESF and ERDF Operational Programmes, 07/10/2022, Brussels (video conference, MTeams)
10. Italian Labour Union (UIL), 10/10/2022, Brussels (video conference, MTeams)
11. Emilia-Romagna region (Italy) – ESF and ERDF Operational Programmes, 10/10/2022, Brussels (video conference, MTeams)

The EU tracks social expenditure for several reasons, ranging from ensuring financial control to assessing whether policies make the desired impact. This study:

- examines the **current state of social tracking** in the Multiannual Financial Framework and the Recovery and Resilience Facility;
 - explores **how it could be improved** to enhance the scrutiny of EU social spending;
 - puts forward a **proposal for a pilot study** of a theory-based monitoring approach as a potential innovation in social tracking.
-