

BENEFITS

Deliverable 2.1

European Green Book for Social Services



Acknowledgments

This report has been developed within the framework of the Horizon Europe project: Building Economic, Needs-Based and Environmental evaluation Frameworks for Inclusive Transformation of Social services in Europe (BENEFITS).

The project partners wish to thank all those who have contributed to the development of this report.

Deliverable Number and Title

D2.1 – European Green Book for Social Services

Authors and Contributors

Leader Developer: **Pinar Cakiroglu** (HEADWAY)

Authors: **Pinar Cakiroglu** (HEADWAY)

Reviewers: **Nikos Avgeris** (EthosLab)

This report should be cited as follows:

Cakiroglu, P. (2026). European Green Book for Social Services (Deliverable D2.1). Athens: BENEFITS Project (Grant Agreement No. 101179032) – Horizon Europe.



Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Transparency Note: The author used Claude (Anthropic) as a writing assistance tool during the drafting of this report. All research design decisions, including the selection and tiering of evaluation methods, the adaptation of HM Treasury Green Book and Vademecum principles to the EU social services context, the development of the three-tier framework (Foundation / Intermediate / Advanced), the design of case study templates, and all analytical and scientific judgements, were made exclusively by the author. AI assistance was limited to structuring and drafting text and templates based on author-provided content, direction, and review.



Funded by
the European Union

Contents

Acknowledgments.....	2
Authors and Contributors	2
Executive Summary	9
INTRODUCTION	11
Practitioner Quick Start Guide.....	14
Quick Start Guide to the European Green Book for Social Services.....	14
CHAPTER 1: A PRACTICAL FRAMEWORK FOR SOCIAL SERVICE EVALUATION.....	18
1.1 Purpose and Scope of this Green Book.....	18
1.2 The Strategic Case for Social Service Evaluation.....	19
1.2.1 The Challenge of Multidimensional Value	19
1.2.2 Temporal Mismatch and Appropriate Time Horizons	20
1.2.3 Non-Linear Causation and Complex Systems	20
1.2.4 Implications for Evaluation Design	21
1.3 Foundations and Essential Elements of Programme Appraisal	22
1.3.1 International Foundations Informing This Framework.....	22
1.3.2 Essential Elements of Programme Appraisal.....	25
1.4 The Proportionality Principle.....	27
1.5 The Three-Tier Methodological Framework	28
1.6 EU Integration and Regulatory Compliance	31
1.7 Core Quality Principles	33
CHAPTER 2: DATA GOVERNANCE AND MANAGEMENT STANDARDS	37
2.1 Data as the Foundation for Evaluation.....	37
2.2 Minimum Viable Data Requirements	37
2.3 Data Planning Standards.....	38
2.4 Data Quality Management Standards.....	39
2.5 GDPR and Data Protection Standards	41
2.6 Data Management Planning.....	43
2.7 Implementation and Resources.....	44
CHAPTER 3: METHODOLOGICAL STANDARDS AND SELECTION	46
3.1 Method Selection Framework	46
3.1.1 How to Use This Chapter	46
3.1.2 Three-Tier Framework Overview.....	47
3.1.3 Method Selection by Primary Evaluation Question	48
3.1.4 Method Selection by Programme Scale	49
3.1.5 Data Requirements by Method.....	49
3.1.6 Combining Methods.....	50
3.1.7 Quality Principles Across All Methods	51
3.1.8 Reading This Chapter: Method Section Structure	51





3.2 FOUNDATION METHODS (Mandatory Standards)	52
3.2.1 Theory of Change	52
3.2.2 Outcome Monitoring	58
3.2.3 Stakeholder Feedback	63
3.3 INTERMEDIATE METHODS (Standards and Commissioning Triggers).....	68
3.3.1 Cost-Effectiveness Analysis	68
3.3.2 Multi-Criteria Decision Analysis	73
3.3.3 Social Return on Investment.....	77
3.4 ADVANCED METHODS (Comprehensive Valuation and Causality)	82
3.4.1 Cost-Benefit Analysis	83
3.4.2 Quasi-Experimental Design	88
3.4.3 Randomised Controlled Trials.....	90
3.4.4 Realist Evaluation	93
3.5 Commissioning External Specialists for Intermediate-Advanced Methods	96
CHAPTER 4: EU INTEGRATION AND REGULATORY COMPLIANCE.....	100
4.1 EU Evaluation Mandates and Method Alignment.....	100
4.2 Data Protection and GDPR Compliance	101
4.3 Cross-Border Harmonisation for Multi-Country Programmes	102
4.4 Reporting Standards for EU Programmes	104
4.5 Reference Standards and Resources	105
CHAPTER 5: PRESENTATION OF RESULTS	107
5.1 Purpose of Results Presentation.....	107
5.2 Core Principles for Results Presentation	107
5.3 Report Structure	108
5.4 Visual Presentation.....	110
5.5 Presenting Uncertainty and Sensitivity Analysis	111
5.6 Reporting Non-Monetised Outcomes.....	111
5.7 Presenting Distributional Effects.....	112
5.8 Reporting for Different Audiences	112
5.9 Quality Standards for Reports.....	113
5.10 Common Reporting Errors	114
CONCLUSIONS.....	115
CASE STUDIES	117
CASE STUDY 1: Theory of Change.....	117
THE CHALLENGE	117
THE INTERVENTION	118
WHY THEORY OF CHANGE WAS SELECTED.....	118
HOW THE THEORY OF CHANGE DEVELOPMENT WAS CONDUCTED.....	118
KEY FINDINGS	119
KEY DATA SUMMARY	120



SOURCES	120
CASE STUDY 2: Cost-Effectiveness Analysis	121
THE CHALLENGE	121
THE INTERVENTION	122
WHY COST-EFFECTIVENESS ANALYSIS WAS CHOSEN.....	122
HOW THE COST-EFFECTIVENESS ANALYSIS WAS CONDUCTED	122
KEY FINDINGS	123
KEY DATA SUMMARY	124
SOURCES	125
CASE STUDY 3: Cost-Benefit Analysis.....	126
THE CHALLENGE	126
THE INTERVENTION	127
WHY COST-BENEFIT ANALYSIS WAS CHOSEN.....	127
HOW THE COST-BENEFIT ANALYSIS WAS CONDUCTED	127
KEY FINDINGS	128
KEY DATA SUMMARY	129
SOURCES	130
CASE STUDY 4: Social Return on Investment (SROI)	131
THE CHALLENGE	131
THE INTERVENTION	131
WHY SOCIAL RETURN ON INVESTMENT WAS CHOSEN.....	132
HOW THE SOCIAL RETURN ON INVESTMENT WAS CONDUCTED.....	132
KEY FINDINGS	133
KEY DATA SUMMARY	134
SOURCES	135
CASE STUDY 5: Mixed Foundation Methods – Process and Outcome Evaluation	136
THE CHALLENGE	136
THE INTERVENTION	137
WHY FOUNDATION METHODS WERE CHOSEN	137
HOW THE FOUNDATION METHODS EVALUATION WAS CONDUCTED.....	137
KEY FINDINGS	138
KEY DATA SUMMARY	140
SOURCES	141
ANNEXES.....	143
ANNEX STRUCTURE	143
ANNEX A.....	147
ANNEX B.....	152
Bibliography.....	251

List of acronyms and tables

Acronyms

Acronym	Full Term
BCR	Benefit-Cost Ratio
CBA	Cost-Benefit Analysis
CEA	Cost-Effectiveness Analysis
CEFR	Common European Framework of Reference for Languages
CMO	Context-Mechanism-Outcome (configuration)
CONSORT	Consolidated Standards of Reporting Trials
CSRD	Corporate Sustainability Reporting Directive
CSQ	Client Satisfaction Questionnaire
DAC	Development Assistance Committee (OECD)
DALY	Disability-Adjusted Life Year
DMP	Data Management Plan
DRG	Diagnosis-Related Group
DWP	Department for Work and Pensions (UK)
EEA	European Economic Area
EQ-5D	EuroQol Five-Dimensional Questionnaire
ESF	European Social Fund
ESF+	European Social Fund Plus
ESOL	English for Speakers of Other Languages
ESRS	European Sustainability Reporting Standards
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable (data principles)
FTE	Full-Time Equivalent
GAD-7	Generalised Anxiety Disorder 7-item scale
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GM	Greater Manchester
GMCA	Greater Manchester Combined Authority
GP	General Practitioner
HACT	Housing Associations' Charitable Trust
HORIZON Europe	EU Framework Programme for Research and Innovation (2021–2027)
ICER	Incremental Cost-Effectiveness Ratio
ICF	ICF Consulting (evaluation firm)
INNOSI	Innovative Social Investment (Horizon 2020 project)
IPS	Individual Placement and Support
IPSE	IPS supplemented with cognitive remediation and social skills training
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISRCTN	International Standard Randomised Controlled Trial Number
JSA	Jobseeker's Allowance (UK)

LSNS	Lubben Social Network Scale
MAXQDA	Software for qualitative and mixed-methods data analysis
MCDA	Multi-Criteria Decision Analysis
MHCLG	Ministry of Housing, Communities and Local Government (UK)
MRC	Medical Research Council (UK)
NEET	Not in Employment, Education, or Training
NEF	New Economics Foundation
NGO	Non-Governmental Organisation
NHS	National Health Service (UK)
NI	National Insurance (UK)
NPS	Net Promoter Score
NPV	Net Present Value
OECD	Organisation for Economic Co-operation and Development
PES	Public Employment Services
PHQ-9	Patient Health Questionnaire 9-item scale
PPP	Purchasing Power Parity
PSSRU	Personal Social Services Research Unit (UK)
QALY	Quality-Adjusted Life Year
QED	Quasi-Experimental Design
RAMESES	Realist And Meta-Narrative Evidence Syntheses: Evolving Standards
RCT	Randomised Controlled Trial
REA	European Research Executive Agency
SAU	Service as Usual
SEK	Swedish Krona
SILC	Survey on Income and Living Conditions (EU-SILC)
SMI	Severe Mental Illness
SROI	Social Return on Investment
ToC	Theory of Change
UCLA	University of California, Los Angeles
UKE	University Medical Centre Hamburg-Eppendorf (Universitätsklinikum Hamburg-Eppendorf)
WEMWBS	Warwick-Edinburgh Mental Wellbeing Scale
WM	West Midlands
WP	Work Package

Tables

Table 1 – Programme Scale Identification	14
Table 2 - Small-Scale Programme Quick Guidance.....	14
Table 3 - Medium-Scale Programme Quick Guidance	15
Table 4 - Large-Scale Programme Quick Guidance	16
Table 5 - Guiding Foundations and Criteria	23
Table 6 - Method Selection by Evaluation Question	48
Table 7 - Data Requirements by Method	49
Table 8 - Theory of Change Programme Snapshot.....	117
Table 9 - Cost Effectiveness Analysis Programme Snapshot.....	121
Table 10 - Cost Benefit Analysis Programme Snapshot.....	126
Table 11 - SROI Analysis Programme Snapshot	131
Table 12 - Mixed Methods Analysis Programme Snapshot	136
Table 13 - Glossary of Key Evaluation Terms.....	143

Templates

Template 1 - Simple Data Management Plan (DMP) Template	147
Template 2 - Data Quality Checklist	150
Template 3 - Cross-Border Data Transfer Checklist	151
Template 4 – Theory of Change Assumption Testing Matrix.....	156
Template 5 – Outcome Monitoring Statement Template.....	167
Template 6 – Outcome Monitoring Quality Assurance Plan	168
Template 7 – Outcome Monitoring Analysis and Reporting Schedule	168
Template 8 – Outcome Monitoring Participant Register Template	175
Template 9 – Outcome Monitoring Outcome Data Template	175
Template 10 – Outcome Monitoring Participation Reporting.....	177
Template 11 – Stakeholder Mapping Matrix	187
Template 12 – CEA Cost Data	202
Template 13 – CEA Outcome Data.....	203
Template 14 - Cost-Effectiveness Ratios	203
Template 15 – CEA Sensitivity Analysis	204
Template 16 - MCDA Weights and Performance Matrix	209
Template 17 – MCDA Sensitivity Analysis A	209
Template 18 – MCDA Sensitivity Analysis B	210
Template 19 - Cross-Border Evaluation Feasibility Assessment by Country.....	239
Template 20 – Cultural Adaptation Documentation	242
Template 21 – Data Dictionary Template.....	244
Template 22 - Cross-Border Partnership Governance	248
Template 23 - Cross-Border Data Protection & Ethics Coordination.....	248
Template 24 – Cross-Border Quality Assurance Coordination	249

Executive Summary

The European Green Book for Social Services (Deliverable D2.1) establishes the first European-level evaluation framework designed specifically for social service interventions. Developed within the BENEFITS (Building Economic, Needs-Based and Environmental evaluation Frameworks for Inclusive Transformation of Social services in Europe) Horizon Europe project, this framework responds to a critical evidence gap: across European social service provision, systematic evaluation of programme effectiveness and value for money remains the exception rather than the norm.

The framework adapts established evaluation methodology — drawing principally on the UK Treasury Green Book and the EU Better Regulation Guidelines — to address three challenges distinctive to social services: multifaceted value that resists monetary reduction, benefits that materialise over time exceeding standard funding cycles, and complex causal pathways operating within dynamic systems.

The Three-Tier Methodological Architecture

The framework is organised around a proportionality principle: evaluation effort must match programme scale, decision stakes, and available resources.

Foundation Methods (Tier 1) — mandatory for all programmes — comprise Theory of Change, Outcome Monitoring, and Stakeholder Feedback. These methods establish the essential building blocks for learning and accountability and can be implemented using internal capacity with modest external support.

Intermediate Methods (Tier 2) — for established programmes facing efficiency questions — include Cost-Effectiveness Analysis, Multi-Criteria Decision Analysis, and Social Return on Investment. These methods require external support and/or expertise in evaluation but remain feasible for medium-scale programmes.

Advanced Methods (Tier 3) — for large-scale programmes or high-stakes policy decisions — encompass Cost-Benefit Analysis, Social Return on Investment, Quasi-Experimental Design, Randomised Controlled Trials, and Realist Evaluation. These methods require specialist expertise and substantial investment.

Practical Guidance and Implementation Support

The framework provides differentiated guidance matched to user needs. The main text (Chapters 1–5) establishes standards and principles. A Practitioner Quick Start Guide directs readers to relevant sections based on programme scale. Five Case Studies demonstrate methods in practice across diverse European social service contexts, drawing on published evaluations from Finland, Denmark, England, and Germany. Comprehensive Annexes provide implementation templates for Foundation and Intermediate Methods and commissioning specifications for Advanced Methods.

EU Integration

The framework aligns with European regulatory requirements including EU Better Regulation Guidelines, European Pillar of Social Rights, GDPR, HORIZON Europe deliverable standards, and ESF+ performance requirements. Dedicated guidance addresses cross-border harmonisation for multi-country programmes, balancing standardisation for comparison with local adaptation respecting national contexts.

Relationship to Other BENEFITS Deliverables

D2.1 provides the conceptual and methodological foundation for subsequent work package deliverables. D1.1 (Meta-Analysis of Existing Indicators) informed the framework's approach to multidimensional value measurement. D2.2 will develop microsimulation modelling tools. D2.3 will pilot-test the framework with social service providers across seven EU Member States, providing empirical validation of the three-tier approach.

INTRODUCTION

The European Union funds social services that reach millions of people across Member States every year as part of employment programmes, mental health support, housing interventions, disability services, youth inclusion initiatives, and community-based care. Yet systematic evidence on whether these services achieve their intended outcomes and at what cost, or what alternatives may serve the needs better remains remarkably thin.

This evaluation gap has economic, social, and political consequences. Programmes that cannot demonstrate their impact do not only become economically and socially unfeasible but also politically vulnerable. When public resources tighten or political priorities shift, services without evidence of effectiveness are the first to lose funding — regardless of their actual value to the people they serve. Evaluation is not merely an academic exercise or a bureaucratic compliance requirement; it is a democratic defence mechanism for social services and the populations they support.

This deliverable, D2.1 of the BENEFITS (Building Economic, Needs-Based and Environmental evaluation Frameworks for Inclusive Transformation of Social services in Europe) project, responds to this gap. It provides a European-level evaluation framework designed specifically for social services, adapting the methodological rigour of the UK Treasury Green Book and the EU Better Regulation Guidelines to the distinctive challenges of social service contexts: multidimensional value that resists monetisation, benefits that materialise over timescales exceeding funding cycles, and complex causal pathways operating within complex, dynamic systems.

The framework is built on a proportionality principle recognising that one of the main reasons of weak systemic evidence on the effectiveness of social services is the lack of capacity in the small to medium-sized organisations providing social services around the EU. Although considered as gold standards in evaluation, not every programme requires a randomised controlled trial or a comprehensive cost-benefit analysis. Carefully conducted simple evaluation creates evidence that is better than no evidence. Small organisations with limited evaluation capacity need practical, implementable methods that produce credible evidence. While large-scale programmes facing high-stakes policy decisions require rigorous causal inference and economic valuation. This Green Book addresses both ends of the spectrum, as well as the needs between, through a three-tier methodological architecture:

Foundation Methods (Theory of Change, Outcome Monitoring, Stakeholder Feedback) establish the essential building blocks every programme can implement regardless of scale. This is the section where we target small to medium-sized organisations to initiate programme evaluation with internal capacity. **Intermediate Methods** (Cost-Effectiveness Analysis, Multi-Criteria Decision Analysis, Social Return on Investment) address efficiency questions and resource allocation decisions for established

programmes. **Advanced Methods** (Cost-Benefit Analysis, Quasi-Experimental Design, Randomised Controlled Trials, Realist Evaluation) provide definitive causal evidence and comprehensive economic valuation for large-scale programmes and high-stakes policy decisions.

Each method is presented with standards and principles in the main text (Chapters 1–5), demonstrated through real-world Case Studies drawing on published evaluations across European social service contexts, and supported by implementation templates or commissioning specifications in the Annexes (A–B.9).

This document should be read alongside other BENEFITS work package deliverables. D1.1 provides the meta-analysis of Beyond-GDP wellbeing indicators that informed the framework's approach to multidimensional value measurement. D2.2 will develop the microsimulation modelling tools for estimating programme impacts at scale. D2.3 will pilot-test the framework with social service providers across seven EU Member States.

Although this framework focuses primarily on the evaluation of existing programmes, the methods presented — particularly Cost-Benefit Analysis (Chapter 3.4.1) and Multi-Criteria Decision Analysis (Chapter 3.3.2) — also serve ex-ante appraisal functions, supporting decision-makers in selecting between policy options before programme launch. The appraisal-evaluation cycle is continuous: ex-ante assessment informs programme design, whilst ongoing and ex-post evaluation feeds back into future appraisal decisions.

How to read this document:

Readers are not expected to read this document cover to cover. The **Practitioner Quick Start Guide** (immediately following this Introduction) provides a pathway based on programme scale, directing readers to the most relevant chapters, case studies, and annexes. **Chapter 1** establishes the conceptual framework and should be read by all users. **Chapter 2** covers data governance and is essential for organisations beginning evaluation for the first time. **Chapter 3** is the methodological core — readers should navigate to sections relevant to their tier as directed at the beginning of the chapter. **Chapters 4–5** address EU regulatory compliance and results presentation respectively. **Case Studies** demonstrate methods in practice. **Annexes** provide templates for implementation (Foundation and Intermediate Methods) and commissioning specifications and quality standards for engaging external evaluators (Advanced Methods). A **Glossary** of key evaluation terms is provided at the beginning of the Annexes for readers unfamiliar with technical terminology.

Practitioner Quick Start Guide



Practitioner Quick Start Guide

Quick Start Guide to the European Green Book for Social Services

This guide helps you navigate the Green Book based on your role and programme context. Read this first, then follow the signposts to the sections and annexes most relevant to your situation.

STEP 1: IDENTIFY YOUR PROGRAMME SCALE

Table 1 – Programme Scale Identification¹

	Small-Scale	Medium-Scale	Large-Scale
Annual budget	Under €500,000	€500,000–€5 million	Over €5 million
Participants per year	Under 500	500-5,000	Over 5,000
Evaluation capacity	No dedicated M&E staff	Some internal M&E capacity	Dedicated evaluation team or budget for external evaluators
Typical decisions	Should we continue? How can we improve?	Which delivery model is most efficient? How do we demonstrate value?	Should this be scaled nationally? Does the evidence justify the investment?

STEP 2: FOLLOW YOUR PATHWAY

PATHWAY A: Small-Scale Programmes

You need: Foundation Methods — the essential building blocks every programme should have.

Table 2 - Small-Scale Programme Quick Guidance

What to do	Where to find it	Time needed
1. Build your Theory of Change — Map how your activities lead to outcomes	Section 3.2.1 (principles) → Annex B.1 (templates and workshop guide)	1–2 days with team

¹ Scale thresholds are indicative. Organisations should apply judgement where programme budgets fall near boundaries, considering also participant numbers and evaluation capacity alongside budget.

2. Set up Outcome Monitoring — Track whether participants' lives improve	Section 3.2.2 (principles) → Annex B.2 (templates, instruments, spreadsheet)	1 week to design, then ongoing
3. Collect Stakeholder Feedback — Hear from participants and staff	Section 3.2.3 (principles) → Annex B.3 (feedback templates)	1 day to design, then quarterly
4. Get your data governance right — GDPR compliance, consent, storage	Chapter 2 (standards) → Annex A.1 (Data Management Plan template)	1–2 days

Essential annexes: A.1, B.1, B.2, B.3 *Skip for now:* B.4–B.9 (these serve larger programmes) *Read if time allows:* Case Study 1 (ToC) and 5 (Foundation Methods) in practice

PATHWAY B: Medium-Scale Programmes

You need: Foundation Methods (if not yet in place) PLUS Intermediate Methods addressing your specific evaluation questions.

Table 3 - Medium-Scale Programme Quick Guidance

What to do	Where to find it	Time needed
1. Ensure Foundation Methods are solid	Pathway A above	—
2. Choose your Intermediate Method(s):		
→ <i>"Which approach is most efficient?"</i>	Section 3.3.1 CEA → Annex B.4 (commissioning specs)	1–2 months (with external support)
→ <i>"How do we weigh competing priorities?"</i>	Section 3.3.2 MCDA → Annex B.5 (facilitation guide)	3–5 days (workshop format)
→ <i>"What social value do we create per € invested?"</i>	Section 3.3.3 SROI → Annex B.6 (commissioning specs)	1–3 months (external SROI practitioner)
3. Commission external support	Section 3.5 (commissioning guidance)	—

Essential annexes: A.1, B.1–B.3 (Foundation), plus B.4, B.5, or B.6 as relevant *Read:* Case Study 2 (CEA) and Case Study 4 (SROI)

PATHWAY C: Large-Scale Programmes / High-Stakes Decisions

You need: Foundation + Intermediate + Advanced Methods. You will be commissioning specialist evaluators.

Table 4 - Large-Scale Programme Quick Guidance

What to do	Where to find it	Time needed
1. Ensure Foundation Methods are solid	Pathway A above	—
2. Choose your Intermediate Method(s):		
→ <i>"Which approach is most efficient?"</i>	Section 3.3.1 CEA → Annex B.4 (commissioning specs)	1–2 months (with external support)
→ <i>"How do we weigh competing priorities?"</i>	Section 3.3.2 MCDA → Annex B.5 (facilitation guide)	3–5 days (workshop format)
→ <i>"What social value do we create per € invested?"</i>	Section 3.3.3 SROI → Annex B.6 (commissioning specs)	1–3 months (external SROI practitioner)
3. Commission external support	Section 3.5 (commissioning guidance)	—

Essential annexes: All — share full document with commissioned evaluators *Read:* Case Study 3 (CBA), Chapter 4 (EU reporting standards)

STEP 3: KEY PRINCIPLES (APPLICABLE TO ALL PATHWAYS)

Whatever your programme scale, these principles apply:

Data governance first: Set up GDPR-compliant data collection before gathering any participant data (Chapter 2, Annex A.1).

Theory of Change is non-negotiable: Every programme needs one. It doesn't have to be elaborate for small programmes — a one-page visual model with 3–5 causal pathways is sufficient (Section 3.2.1).

Start where you are: If your programme currently collects no evaluation data, begin with Foundation Methods. Perfect data is the enemy of any data. Implementing basic outcome monitoring is a significant step forward.

Proportionality, not perfection: Match evaluation effort to your programme's scale and the decisions it needs to inform. A well-implemented Foundation evaluation is more valuable than a poorly resourced Advanced evaluation.

Evaluation is ongoing: These are not one-time exercises. Theory of Change should be reviewed annually. Outcome monitoring is continuous. Stakeholder feedback is collected quarterly at minimum.

Chapter 1:

A practical framework for social service evaluation



CHAPTER 1: A PRACTICAL FRAMEWORK FOR SOCIAL SERVICE EVALUATION

1.1 Purpose and Scope of this Green Book

The European Green Book for Social Services provides definitive guidance for appraising and evaluating social service interventions within the European Union. Its purpose is to ensure public resource allocation is supported by objective analysis that maximises social value across Member States.

This guidance applies to public servants, practitioners, policy makers, and researchers involved in designing, funding, delivering, or evaluating social programmes and projects. It aims to establish methodological standards for assessing social welfare costs, benefits, and trade-offs of implementation options. This Green Book provides a framework for accountability, rigour, and strategic alignment in policy development—not a mechanical decision-making device.

Social value refers to total welfare and wellbeing generated for populations served. This extends beyond financial savings to include non-market impacts: dignity, autonomy, community cohesion, social inclusion, and other outcomes often omitted from traditional financial appraisals. Social value encompasses both monetised benefits (earnings, cost savings, resource use) and non-monetised benefits that resist monetary valuation but remain critical to decision-making.

Economic versus financial analysis: This Green Book focuses on economic appraisal—assessing social value through costs and benefits to society regardless of who bears them. Economic analysis differs from financial analysis, which assesses cash flows and budget impacts for implementing organisations. Both analyses are required for comprehensive decision-making:

Economic analysis, as covered here, measures economic and social value: welfare to society, costs and benefits regardless of who bears them, expressed in discounted real terms (inflation-adjusted), answers "does this create more value than it costs?"

Financial analysis measures budget impacts: cash flows to implementing organisations, affordability within budget constraints, expressed in nominal terms, answers "can we afford this?"

This Green Book provides methods for economic analysis. Financial analysis follows standard budgeting and accounting practices detailed in organisational financial management guidance, not covered here.

1.2 The Strategic Case for Social Service Evaluation

Social services differ fundamentally from infrastructure and regulatory interventions traditionally addressed by evaluation frameworks like the UK Treasury Green Book or EU Cost-Benefit Analysis guidelines. Whilst these frameworks provide valuable foundations, they were designed primarily for contexts with measurable physical outputs, clear causal pathways, and market-determined values. Social services present three fundamental evaluation challenges that require adapted methodological approaches: the challenge of multidimensional value, temporal complexity, and non-linear causation.

1.2.1 The Challenge of Multidimensional Value

Social programmes generate outcomes that cannot be fully captured through market mechanisms or monetary valuation alone. Dignity, justice, community cohesion, family stability, personal autonomy, voice, empowerment — these outcomes are real, consequential, and central to programme success, yet resist straightforward monetisation. This recognition is grounded in extensive scholarship: Sen's capability approach establishes that human wellbeing comprises functionings and freedoms irreducible to income or utility (Sen, 1999); the Stiglitz-Sen-Fitoussi Commission demonstrated that GDP and monetary metrics systematically fail to capture dimensions of wellbeing essential for policy evaluation (Stiglitz, Sen and Fitoussi, 2009); and Nussbaum's capabilities framework identifies central human capabilities — including bodily integrity, affiliation, and control over one's environment — that constitute the proper objectives of social policy yet resist market valuation (Nussbaum, 2011). Forcing all outcomes into monetary terms can produce arbitrary valuations that exclude critical benefits from decision-making or reduce genuine complexity to misleading simplicity (Ackerman and Heinzerling, 2004).

These foundational critiques continue to shape contemporary measurement practice. A meta-analysis of 66 Beyond-GDP indicators conducted within the BENEFITS project (D1.1) confirms that multidimensional wellbeing frameworks have proliferated dramatically since the Stiglitz-Sen-Fitoussi Commission report, with 65% of existing frameworks developed after 2010. Yet the same analysis reveals persistent blind spots directly relevant to this Green Book: care services appear in only 20% of indicators, service integration in 11%, and prevention in 6% — precisely the domains where social service evaluation must operate (Kubiszewski et al., 2025; Tzivanakis, Melios et al., 2025)

Rigorous evaluation must accommodate both monetised and non-monetised value. Decision-makers require transparency about what has been monetised, which methods were used, what remains unmonetised and why, and what trade-offs exist between competing values. Summary metrics that reduce all outcomes to a single number can obscure these critical trade-offs rather than illuminate them.



This Green Book addresses multidimensional value through methods that make trade-offs explicit. Cost-Benefit Analysis (3.4.1) provides frameworks for monetisation where appropriate and defensible. Multi-Criteria Decision Analysis (3.3.2) enables systematic comparison across diverse outcome dimensions without forcing monetary conversion. Cost-Effectiveness Analysis (3.3.1) allows comparison of programmes achieving similar goals using different approaches. Together, these methods enable decision-makers to understand programme value comprehensively rather than forcing reduction to single metrics.

1.2.2 Temporal Mismatch and Appropriate Time Horizons

Returns on investment in social services often materialise over timelines exceeding standard funding cycles and evaluation windows. Early intervention with children, preventative health programmes, education and skills development, community infrastructure building—benefits may not emerge for years or decades. Evaluations conducted within short timeframes risk concluding programmes are ineffective when long-term benefits have not yet matured: improved life chances, prevented chronic conditions, reduced long-term service use, intergenerational effects.

Evaluation design must account for appropriate time horizons even when these exceed funder reporting requirements. This requires distinguishing between:

- Immediate outputs: Services delivered, participants reached, activities completed
- Short-term outcomes: Skills acquired, immediate wellbeing improvements, behaviour changes
- Medium-term outcomes: Sustained behaviour change, service use reduction, employment or education participation
- Long-term impacts: Life course changes, prevented adverse events, societal-level effects, intergenerational benefits

Chapter 3 provides detailed guidance on selecting appropriate evaluation timeframes matched to programme logic and expected benefit timelines. Theory of Change (3.2.1) enables specification of expected causal pathways with anticipated timeframes for different outcome categories. Whilst practical constraints may limit evaluation duration, transparency about anticipated but unmeasured long-term benefits ensures decision-makers understand programme value comprehensively rather than confusing absence of evidence with evidence of absence.

1.2.3 Non-Linear Causation and Complex Systems

Social interventions operate as complex human interactions within dynamic systems, not simple linear "if-then" mechanisms. Programme success depends on multiple interacting factors: local context (labour markets, complementary services, institutional capacity), relationships (trust between staff and participants, peer dynamics, family





support), implementation quality (staff skills, fidelity to programme model, appropriate adaptation to local need), complementary services (health, housing, employment support operating simultaneously), external factors (economic conditions, policy changes, community resources), and interactions between these elements that produce emergent effects unpredictable from individual components alone.

Simple before-after comparisons or attribution to single programmes often misrepresent this reality. Participants typically receive multiple services simultaneously. Programmes operate within complex service systems where boundaries are porous. External factors affect outcomes independent of programme activities. Changes frequently result from combinations of factors rather than single interventions producing single outcomes through single pathways.

Evaluation methods must accommodate this complexity whilst maintaining analytical rigour. Theory of Change (3.2.1) makes causal assumptions explicit and testable. Realist Evaluation approaches (3.4.4) examine what works, for whom, in what contexts, and why. Quasi-experimental design and randomised control trials (3.4.3 and 3.4.4) address selection bias and confounding whilst acknowledging limitations of causal inference in complex systems. These methods enable robust conclusions about programme contribution to observed changes whilst avoiding false certainty about isolated causal effects.

Consider a housing support programme for formerly homeless individuals. Success depends not only on housing quality but also on availability of employment opportunities, mental health services, peer support networks, staff-client relationships, local transport infrastructure, welfare benefit systems, and participants' own agency and social networks. Evaluation approaches that ignore this complexity—attributing all outcomes to housing alone—miss the reality that housing is necessary but insufficient, that effectiveness depends on complementary factors, and that similar programmes may succeed in one context and fail in another based on these contextual differences.

1.2.4 Implications for Evaluation Design

These three challenges are not obstacles to rigorous evaluation—they are design requirements. The methodological framework presented in this Green Book addresses each challenge directly:

Multidimensional value: Methods accommodate both monetised and non-monetised outcomes, making trade-offs transparent rather than hidden

Temporal complexity: Guidance on appropriate timeframes and proxy indicators enables evaluation even when ultimate impacts require decades to materialise

Non-linear causation: Theory-based and realist approaches make causal assumptions explicit whilst acknowledging inherent complexity

Section 1.3 establishes the foundations upon which this framework builds its response to these challenges.

1.3 Foundations and Essential Elements of Programme Appraisal

1.3.1 International Foundations Informing This Framework

The methodological framework presented in this Green Book adapts established evaluation methodology to the specific challenges of social service contexts identified in Section 1.2. Two primary bodies of guidance provide the structural and technical foundations; a third served as an influential reference confirming the comprehensiveness of the resulting framework.

The UK Treasury Green Book (HM Treasury, 2022) provides the primary methodological foundation. The appraisal discipline that structures this entire document — the distinction between economic analysis (social value to society) and financial analysis (budget impact to organisations), the requirement for strategic justification and SMART objectives before evaluation design begins, the insistence on discounting future values to present terms, the discipline of sensitivity analysis, and the proportionality principle scaling analytical effort to decision significance — derives from the Treasury Green Book. Its proportionality principle directly shapes the three-tier methodology presented in Section 1.5: Foundation Methods for all programmes, Intermediate Methods where efficiency questions justify additional investment, and Advanced Methods where high-stakes decisions require definitive evidence.

However, the Treasury Green Book was designed primarily for infrastructure, regulatory, and fiscal interventions where outputs are physical, causal pathways are relatively linear, and values are largely market-determined. Section 1.2 demonstrates that social services present fundamentally different evaluation challenges. This Green Book therefore adapts Treasury methodology rather than applying it wholesale, extending its appraisal discipline into domains — Theory of Change, stakeholder-driven valuation, realist evaluation, complexity-aware causal methods — that the Treasury framework acknowledges but does not develop for social service contexts.

The EU Better Regulation Guidelines and Cost-Benefit Analysis Vademecum (European Commission, 2021) provide the regulatory and operational context within which this framework operates. For programmes funded under HORIZON Europe, ESF+, and other EU instruments, these guidelines establish mandatory evaluation standards: a 3% social discount rate for economic evaluations, comprehensive benefit identification from a societal perspective, distributional analysis assessing impacts on

vulnerable groups, and stakeholder consultation throughout the evaluation cycle. The EU framework also introduces requirements specific to the European context — cross-border harmonisation, alignment with the European Pillar of Social Rights, GDPR-compliant data governance — that the Treasury Green Book does not address. Where UK Treasury and EU standards diverge, most notably on discount rates (3.5% UK versus 3% EU), this framework defaults to EU standards for EU-funded evaluations and notes UK standards for sensitivity analysis. Chapter 2 operationalises the data governance requirements; Chapter 4 addresses cross-border harmonisation and EU reporting standards.

The OECD DAC Evaluation Criteria (OECD DAC Network on Development Evaluation, 2019; OECD, 2021) influenced the structural and methodological decisions of this framework and served as an international benchmark confirming its comprehensiveness. The six criteria — relevance, coherence, effectiveness, efficiency, impact, and sustainability — represent the most widely adopted evaluation standards globally, defining the normative questions any credible evaluation must address. Originally established in 1991 and revised in 2019 to include the criterion of coherence, the DAC criteria define *what* evaluation should examine but do not prescribe specific methods. During the development of this Green Book, the DAC framework informed method selection decisions — ensuring, for example, that the framework addresses sustainability (whether benefits last beyond funding periods) and coherence (whether programmes fit with other interventions in the same context), dimensions that might otherwise have received insufficient attention. Table 1.1 demonstrates that the methodology developed in this framework addresses all six DAC criteria, providing international validation that no significant evaluation dimension has been omitted.

Table 5 - Guiding Foundations and Criteria

Green Book Appraisal Element	Primary Foundations (HM Treasury / EU Better Regulation)	OECD DAC Criteria Addressed	Green Book Operationalisation
Strategic justification — why the intervention is needed	HM Treasury strategic case: evidence of need, policy alignment, stakeholder consultation	<i>Relevance</i> : Is the intervention responding to genuine needs and priorities?	Theory of Change (3.2.1); Stakeholder Feedback (3.2.3)
SMART objectives — what the programme aims to achieve	HM Treasury: Specific, Measurable, Achievable, Relevant, Time-bound objectives	<i>Relevance and Effectiveness</i> : Are objectives appropriate and being achieved?	Theory of Change (3.2.1); Outcome Monitoring (3.2.2)

Logical foundation — how activities produce outcomes	HM Treasury: Theory of Change / logic model requirement; EU: intervention logic	<i>Effectiveness:</i> Achievement along the results chain	Theory of Change (3.2.1) — mandatory for all programmes
Policy coherence — fit with other interventions	EU Better Regulation: alignment with EU policy frameworks	<i>Coherence:</i> Compatibility with other interventions, internally and externally	MCDA (3.3.2); Chapter 4 cross-border harmonisation
Economic analysis — social value created	HM Treasury: economic appraisal, discounting, sensitivity analysis; EU Vademecum: 3% social discount rate, distributional analysis	<i>Efficiency and Impact:</i> Are resources used well? What broader difference is made?	CEA (3.3.1); SROI (3.3.3); CBA (3.4.1)
Financial feasibility — affordability and budget impact	HM Treasury: financial case, cash flow, affordability	<i>Efficiency:</i> Timely and economic delivery	Financial analysis (distinct from economic analysis — see Section 1.1)
Stakeholder engagement — meaningful participation	HM Treasury: stakeholder consultation; EU Better Regulation: mandatory consultation	<i>Relevance:</i> Responsiveness to those served	Stakeholder Feedback (3.2.3); SROI stakeholder process (3.3.3); MCDA facilitation (3.3.2)
Causal evidence — whether the programme caused observed changes	HM Treasury: impact evaluation; EU Better Regulation: impact assessment for major policies	<i>Effectiveness and Impact:</i> Attribution and contribution	QED (3.4.2); RCT (3.4.3); Realist Evaluation (3.4.4)
Sustainability — whether benefits last	OECD DAC: sustainability criterion (financial, institutional, social dimensions)	<i>Sustainability:</i> Do net benefits continue?	Theory of Change assumption testing; extended outcome monitoring; stakeholder ownership assessment
Implementation capability — capacity to deliver	HM Treasury: management case, governance, risk, delivery arrangements	Implicit across all criteria	Chapter 2 data governance; Annex A implementation templates

1.3.2 Essential Elements of Programme Appraisal

Drawing on these foundations, all social service interventions should address the following elements proportionate to their scale. These elements establish what every programme must consider; the proportionality principle (Section 1.4) governs how deeply each is pursued, and the three-tier methodology (Section 1.5) provides the specific methods.

Strategic justification: Clear statement of why intervention is needed, including evidence of problem magnitude and causes, alignment with policy frameworks (European Pillar of Social Rights, national priorities), stakeholder consultation findings informing design, and identification of intended beneficiaries. Evaluation should assess whether the strategic case remains valid as circumstances evolve.

SMART objectives: Specific, Measurable, Achievable, Relevant, Time-bound objectives defining success (HM Treasury, 2022). For social service contexts, this requires particular attention to feasibility: objectives must be specific enough to guide evaluation design, measurable through feasible data collection methods, achievable given programme resources and implementation timeline, relevant to strategic priorities, and time-bound with clear target dates. Vague objectives ("improve wellbeing," "strengthen communities") without measurable specifications prevent effective evaluation.

Logical foundation: Theory of Change mapping how inputs and activities produce outputs, outcomes, and impacts through causal pathways (see *Glossary*). Explicit statement of assumptions underpinning each causal link. Identification of critical success factors and external dependencies. Theory of Change is essential for all programmes (Section 3.2.1) and forms the foundation for all subsequent evaluation.

Policy coherence: Assessment of how the programme fits with other interventions operating in the same context — both internally (within the implementing organisation or government) and externally (with other actors' efforts serving the same population). Programmes operating in isolation from complementary services, or duplicating existing provision, reduce collective impact. Policy coherence was elevated as a named appraisal element in this framework following the OECD DAC's 2019 addition of coherence as a formal evaluation criterion.

Economic analysis: Assessment of social value created — costs and benefits to society regardless of who bears them. Economic analysis takes a societal perspective, measuring welfare improvements through monetised benefits (earnings, cost savings, resource use) and non-monetised benefits (dignity, autonomy, outcomes resisting valuation). Economic analysis employs discounting to present value, uses real terms (inflation-adjusted), and assesses whether total benefits exceed total costs (HM Treasury, 2022). The EU Better Regulation Vademecum specifies a 3% social discount rate for EU-funded evaluations. This Green Book presents four economic analysis



methods scaled to programme needs: Cost-Effectiveness Analysis (Section 3.3.1), Multi-Criteria Decision Analysis (Section 3.3.2), Social Return on Investment (Section 3.3.3), and Cost-Benefit Analysis (Section 3.4.1).

Financial feasibility: Assessment of budget affordability and cash flow requirements from the implementing organisation's perspective. Financial analysis uses nominal terms, maps expenditure and revenue timing, confirms funding sources, and assesses organisational capacity to manage cash flows. Financial analysis is distinct from economic analysis: programmes creating substantial social value may impose cash requirements exceeding organisational budgets, whilst financially affordable programmes may create minimal social value. Both perspectives are required for decision-making.

Stakeholder engagement: Systematic involvement of participants, families, communities, and delivery partners in programme design and evaluation. Co-production arrangements ensuring meaningful participation in decisions affecting service users. Stakeholder feedback mechanisms capturing participant voice (Section 3.2.3). Ethical governance frameworks protecting vulnerable populations whilst respecting autonomy and dignity. Stakeholder engagement is both a standalone element and a crosscutting principle — it is central to SROI (Section 3.3.3), integral to MCDA facilitation (Section 3.3.2), and essential for Theory of Change development (Section 3.2.1).

Causal evidence: Assessment of whether the programme caused observed changes, as distinct from changes attributable to other factors. The strength of causal evidence required should be proportionate to the decisions the evaluation will inform. Foundation Methods (Section 3.2) provide associational evidence — outcomes improved during programme engagement. Intermediate and Advanced Methods provide progressively stronger causal evidence through comparison group designs (Sections 3.4.2–3.4.3) or theory-based causal analysis (Section 3.4.4).

Sustainability assessment: Consideration of whether programme benefits will continue beyond the current funding period, including financial sustainability (can the programme sustain itself?), institutional sustainability (do implementing organisations have capacity and commitment to continue?), and social sustainability (do participants retain gains over time?). Sustainability is often the weakest element in social service evaluation — programmes demonstrate short-term outcomes but rarely assess whether benefits persist. This element was strengthened in this framework in response to the OECD DAC sustainability criterion, which requires evaluation to examine whether net benefits continue across financial, institutional, and social dimensions.

Implementation capability: Delivery arrangements including governance structures, management systems, staff capacity, risk management, and monitoring systems. Quality assurance processes ensuring fidelity to programme model. Mechanisms for learning and adaptation based on implementation experience and evaluation findings.

Without adequate implementation capability, well-designed programmes fail to produce intended results.

Small programmes address these elements through Foundation Methods (Section 3.2) implemented with internal capacity and modest external support. Larger programmes require comprehensive treatment adding Intermediate Methods (Section 3.3) and Advanced Methods (Section 3.4) with specialist input. Proportionality (Section 1.4) applies to depth of analysis, not to whether elements are addressed.

1.4 The Proportionality Principle

This Green Book establishes a core principle: evaluation effort must be proportionate to programme scale, budget, and timeline. Larger programmes with greater societal stakes warrant more sophisticated methods and external expertise; smaller programmes focus on essential learning using internal capacity. Disproportionate evaluation—either insufficient scrutiny of major programmes or excessive evaluation of small pilots—misallocates analytical resources and produces misleading conclusions. Therefore, a three-tiered method coverage has been implemented:

Foundation level (Tier 1): Every programme regardless of scale should implement essential building blocks for learning and accountability. Theory of Change (causal logic), outcome monitoring (tracking change), and stakeholder feedback (participant voice) are essential minimum standards for credible evaluation. Most programmes can implement these methods using internal capacity, with targeted external support for specific technical elements.

Intermediate level (Tier 2): Programmes with stable implementation facing efficiency questions or complex resource allocation decisions should employ Intermediate Methods. Cost-effectiveness analysis compares alternative delivery approaches. Multi-criteria decision analysis navigates trade-offs between competing objectives. These methods typically require external evaluation support (academic partnerships, specialist consultants) but remain feasible for medium-scale programmes.

Advanced level (Tier 3): Large-scale programmes, high-stakes policy decisions, or programmes requiring definitive proof of impact should employ Advanced Methods. Cost-benefit analysis provides comprehensive economic assessment. Quasi-experimental designs and randomised controlled trials establish causation. Realist evaluation addresses complexity. These methods require specialist expertise (econometric analysis, trial design, advanced statistics) and substantial evaluation investment varying by method and context.



Proportionality applies both to method selection and to analytical depth within methods. Large programmes conducting cost-benefit analysis require comprehensive benefit quantification, sophisticated sensitivity analysis, and independent quality assurance. Small programmes employing cost-effectiveness analysis for internal decision-making require only basic outcome costing and simple comparisons. Analytical effort must match decision stakes.

The proportionality principle does not excuse poor-quality implementation of Foundation Methods. Small programmes must implement Theory of Change, outcome monitoring, and stakeholder feedback rigorously even when analytical sophistication remains limited. Quality standards apply at all scales (Chapter 3.2); what varies by scale is method complexity and external expertise requirements.

1.5 The Three-Tier Methodological Framework

This Green Book organises evaluation methods into three tiers reflecting programme scale, decision stakes, and analytical requirements. The framework ensures proportionate evaluation whilst maintaining quality standards at all levels.

Tier 1: Foundation Methods (Mandatory Standards)

Purpose: Essential building blocks for learning and accountability, required for all programmes regardless of scale.

Methods:

- **Theory of Change (Chapter 3.2.1):** Explicit causal logic mapping how programme activities produce intended outcomes through specified mechanisms. Identifies assumptions, critical success factors, and external dependencies.
- **Outcome Monitoring (Chapter 3.2.2):** Systematic tracking of participant outcomes using validated measures where available or fit-for-purpose indicators where validated instruments do not exist.
- **Stakeholder Feedback (Chapter 3.2.3):** Structured collection and analysis of participant voice, staff perspectives, and partner feedback. Mechanisms for "closing the loop"—demonstrating how feedback informs programme adaptation.

Implementation: Foundation Methods implemented using internal capacity with external support for method design where needed. Timeline: ongoing throughout programme operation, not one-time activities.

Who implements: Programme staff with evaluation skills, internal monitoring and evaluation teams, or external evaluation partners providing training and quality assurance whilst programme staff conduct data collection.

Tier 2: Intermediate Methods (Efficiency, Choice, and Social Value)

Purpose: Methods proving efficiency or supporting complex resource allocation decisions for established programmes with demonstrated implementation capability.

Methods:

- **Cost-Effectiveness Analysis (Chapter 3.3.1):** Compares costs per unit outcome across alternative delivery approaches. Answers "which approach achieves outcome X most efficiently?"
- **Multi-Criteria Decision Analysis (Chapter 3.3.2):** Structures complex decisions involving trade-offs between competing objectives, stakeholder values, and uncertain outcomes. Makes decision logic transparent and defensible.
- **Social Return on Investment (Chapter 3.3.3):** Stakeholder-driven framework for measuring social, environmental, and economic value creation. Quantifies social value per unit invested through participatory outcome identification and financial proxy valuation. Common in social investment contexts (impact bonds, venture philanthropy) where funders require standardised value metrics. Requires accredited practitioners or consultants with expertise in stakeholder engagement and social value measurement. More comprehensive than cost-effectiveness analysis but more accessible than full cost-benefit analysis.

Implementation: Intermediate Methods typically require external evaluation support (academic partnerships, specialist consultancies). Timeline: 6-12 months for complete evaluation cycle.

Who implements: Evaluation specialists with expertise in economics (cost-effectiveness analysis), decision science (multi-criteria analysis), or social value measurement (accredited practitioners). Programme staff provide programme data, facilitate stakeholder engagement, and ensure practical relevance.

Proportionality threshold: Appropriate for established programmes with stable implementation where efficiency gains or allocation decisions justify evaluation investment. Not proportionate for small pilots still refining basic implementation.

Tier 3: Advanced Methods (Comprehensive Valuation and Causality)

Purpose: Rigorous methods providing definitive proof of causal impact or comprehensive economic valuation for large-scale programmes or high-stakes policy decisions.

Methods:

Cost-Benefit Analysis (Chapter 3.4.1): Comprehensive economic evaluation quantifying all costs and benefits in monetary terms. CBA is the gold standard for major policy decisions, forming the core of the UK Treasury Green Book and EU Better Regulation Guidelines. Requires health economists or public policy analysts with demonstrated CBA expertise. Answers "does this programme create more value than it costs to society?" through systematic identification, quantification, and monetisation of all significant impacts from societal perspective.

Quasi-Experimental Design (Chapter 3.4.2): Uses comparison groups and statistical methods to establish whether programmes cause observed outcomes. Employs propensity score matching, difference-in-differences, regression discontinuity, or instrumental variables to control for selection bias. Requires causal inference specialists (econometricians, quantitative social scientists). More feasible than randomised controlled trials but carries residual uncertainty about unmeasured confounding. Answers "did this programme cause these outcomes or would they have occurred anyway?"

Randomised Controlled Trials (Chapter 3.4.3): Random assignment creates equivalent treatment and control groups, eliminating selection bias. Gold standard for causal inference. Requires trial methodologists, statisticians, and research infrastructure (ethics approval, trial registration, data monitoring). Substantial resource requirements and ethical considerations (is randomisation acceptable given programme context?). Answers definitively whether programmes cause observed outcomes.

Realist Evaluation (Chapter 3.4.4): Understands how programmes work in complex contexts by identifying context-mechanism-outcome configurations. Incorporates Contribution Analysis for assessing causation in complex multi-actor environments where attribution to single programmes is impossible. Appropriate when programme success depends on context and implementation variation, or for systems-level interventions and policy changes. Requires skilled qualitative researchers with expertise in complexity methods. Answers "how, why, for whom, and in what contexts does this programme produce effects?"

Implementation: Advanced Methods require specialist expertise and substantial investment varying by method and context. Programme staff facilitate access to data and stakeholders but do not conduct analyses requiring specialist technical expertise.



Who implements: Academic research teams, government analytical services, specialist evaluation consultancies with demonstrated expertise in respective methods.

Proportionality threshold: Appropriate for large-scale programmes, policy decisions affecting substantial populations, scaling decisions requiring definitive impact evidence, or social investment contexts requiring economic returns demonstration. Not proportionate for small-medium programmes where evaluation investment exceeds potential learning value or decision stakes.

Reality check: Most small-medium organisations cannot afford Advanced Methods unless: (1) part of multi-site evaluation sharing costs across sites, (2) funder provides dedicated evaluation grant, or (3) pro-bono academic partnership secured. Organisations lacking these conditions should focus on excellent implementation of Foundation and Intermediate Methods rather than under-resourced Advanced Methods producing unreliable results.

Commissioning Advanced Methods: Chapter 3.5 provides specifications for commissioning external specialists. Technical specifications for implementing Advanced Methods appear in Specialist Toolkit annexes (B.6-B.8) for reference by commissioned specialists, not for implementation by programme staff.

1.6 EU Integration and Regulatory Compliance

This framework aligns with European regulatory requirements and funding mandates whilst respecting subsidiarity and national competencies. Evaluations conducted under this framework should comply with:

EU Better Regulation Guidelines: The EU Better Regulation framework informing this Green Book is established in Section 1.3. Specific compliance requirements for EU-funded programmes include: stakeholder consultation, options analysis, proportionality assessment, and presentation of costs, benefits, and distributional impacts. Cost-benefit analysis follows EU Better Regulation standards for economic evaluation, including 3% social discount rate and comprehensive benefit identification.

European Pillar of Social Rights: Interventions must demonstrate contribution to one or more Pillar principles (equal opportunities and access to labour market; fair working conditions; social protection and inclusion) (European Parliament et al., 2017). Evaluation design should track outcomes aligned with Pillar objectives whilst acknowledging Member State competence for social policy design and implementation.

GDPR and Data Protection: All data collection must comply with GDPR requirements for lawful processing, data minimisation, and special protections for sensitive personal data (European Union, 2016). Cross-border data transfers outside the EU require appropriate safeguards. Chapter 2 provides comprehensive data governance standards



ensuring GDPR compliance and Chapter 4 gives reporting standards and protocols for EU integration.

European Sustainability Reporting Standards (ESRS): Social service providers operating within or alongside organisations subject to the Corporate Sustainability Reporting Directive (CSRD) should be aware of alignment opportunities between evaluation conducted under this Green Book and ESRS reporting requirements (European Commission Delegated Regulation (EU) 2023/2772). Whilst ESRS primarily targets corporate entities, the social impact measurement requirements within ESRS S1 (Own Workforce), S2 (Workers in the Value Chain), S3 (Affected Communities), and S4 (Consumers and End-Users) share substantial methodological common ground with this Green Book's outcome measurement and stakeholder engagement frameworks. Organisations required to report under ESRS may find that Foundation Methods data (particularly Outcome Monitoring and Stakeholder Feedback) can serve dual purposes — satisfying both programme evaluation requirements and ESRS social impact disclosure obligations, reducing duplicative data collection effort.

HORIZON Europe Requirements: For innovation actions and research projects funded under HORIZON Europe, evaluation must meet deliverable standards specified in Grant Agreements including Theory of Change documentation, indicator frameworks, data management plans, open access to evaluation data where feasible, and compliance with Responsible Research and Innovation principles. Multi-site projects require coordination protocols harmonising core indicators whilst allowing local adaptation (Chapter 4).

European Social Fund Plus (ESF+) Performance Framework: For programmes co-financed through European Social Fund Plus, evaluation must track common output indicators (participants reached, service intensity) and result indicators (employment, education, social inclusion outcomes) defined in ESF+ Regulation (Chapter 4) (European Union, 2021).

Principle of Subsidiarity: EU funding and regulatory requirements establish minimum standards. Member States retain competence for social policy design, service delivery models, and evaluation requirements beyond EU minimums. This Green Book provides frameworks applicable across diverse national contexts whilst acknowledging legitimate variation in implementation reflecting constitutional arrangements, administrative traditions, and policy priorities.

Cross-border harmonisation: Multi-country programmes must balance standardisation (enabling comparison and aggregation) with local adaptation (respecting context and competencies). Core indicators should be harmonised across sites; supplementary indicators may vary by country. Chapter 4 provides protocols for cross-border evaluation coordination ensuring comparability whilst maintaining proportionality.

1.7 Core Quality Principles

Regardless of method or tier, all evaluations must adhere to core quality principles that ensure credibility, ethical conduct, and usefulness. Quality principles are not optional extras—they are essential requirements distinguishing rigorous evaluation from superficial monitoring. The following standards apply across all three tiers:

Transparency: All assumptions, data sources, analytical choices, and limitations must be documented explicitly. Methods sections should enable replication by independent analysts given access to the same data. Uncertainty ranges should be reported alongside point estimates. Sensitivity analyses should test robustness of conclusions to analytical choices. Divergent stakeholder perspectives should be reported, not suppressed. Evaluation reports should distinguish what evidence shows from what analysts infer or recommend.

Stakeholder engagement: Participants, families, communities, and delivery partners must be involved meaningfully in evaluation design, data interpretation, and use of findings. Stakeholder engagement serves multiple purposes: ensures evaluation questions reflect stakeholder priorities, incorporates implementation knowledge into evaluation design, validates findings against lived experience, builds ownership of findings, and facilitates use of evaluation for programme improvement.

Ethical conduct: Evaluation must protect vulnerable populations whilst respecting autonomy and dignity. Informed consent obtained before data collection, with clear explanation of purposes, risks, and participants' rights to withdraw. Data managed securely with appropriate access controls. Confidentiality protected in reporting. Evaluation should not impose excessive burden on participants or extract data without reciprocal benefit. Evaluations involving children, people with cognitive impairments, refugees, or other vulnerable groups require heightened ethical protections through research ethics review.

Technical rigour: Methods must be implemented according to established standards. Statistical analyses should follow current practice in relevant discipline. Sample sizes should be adequate for intended inferences. Data quality should be assessed and limitations acknowledged. Causal claims should be justified by appropriate research designs. Valuation methods should be documented and defensible. Evaluation specialists should have demonstrated competence in methods employed—credentials, publications, and prior evaluation experience relevant to methods and populations. Quality assurance should be independent of those with vested interests in evaluation conclusions.



Proportionate quality: Quality standards apply at all tiers but implementation varies by scale. Small programmes conducting outcome monitoring require validated measures or well-justified fit-for-purpose indicators, adequate follow-up rates (>70%)², and clear documentation—not sophisticated statistical models. Large programmes conducting cost-benefit analysis require comprehensive benefit quantification, formal sensitivity analysis, and independent quality assurance. Proportionality means matching analytical sophistication to decision stakes, not excusing poor implementation of basic quality standards.

Independence: Evaluators should be independent from programme management to enable objective assessment. Internal evaluations conducted by programme staff face inherent conflicts of interest and should be supplemented by external evaluation for medium-large programmes. For small programmes where external evaluation is not feasible, internal evaluators should report to governance boards rather than programme managers, and findings should be subject to critical review by external peers or funders. Independence does not require evaluator scepticism toward programmes—developmental evaluation approaches that support programme improvement whilst maintaining objectivity often prove more useful than purely summative audit. However, evaluators must have no personal stake in evaluation conclusions beyond professional reputation for honest reporting.

Utility: Evaluations should be designed to inform decisions, not merely document activities. Evaluation questions should reflect real decision needs. Timing should align with decision points. Reporting should communicate to decision-makers in accessible language whilst maintaining technical precision. Evaluations producing technically sophisticated analyses that arrive too late, address irrelevant questions, or remain impenetrable to non-specialists fail the utility test regardless of methodological rigour. Evaluators should engage with decision-makers throughout the evaluation process, not only at final report delivery, to ensure evaluation remains relevant as contexts evolve.

Evaluation should be designed alongside programme development, not retrofitted after implementation begins. Data collection systems, baseline measurement, and Theory of Change should be established before or at programme launch. In some cases, ex-ante appraisal precedes programme design, and ex-post evaluation may continue well beyond programme completion where long-term impacts require extended follow-up.

These principles apply across all three tiers. Quality principles are not optional extras but essential requirements for credible evaluation. Subsequent chapters detail how these principles apply to specific methods: Chapter 2 provides data governance protocols ensuring ethical conduct and GDPR compliance; Chapter 3 specifies quality standards

² See Practitioner Alert Box: When 70% Follow-Up Is Difficult, Section 2.3.

for each methodological tier; Annexes provide evaluation templates embedding quality principles into evaluation design.

How to Use This Green Book

This guidance is organised to support readers at different programme scales:

All programmes: Read Chapters 1-2 (framework and data governance) and implement Foundation Methods (Chapter 3.2)

Medium-scale programmes with efficiency questions: Add Intermediate Methods (Chapter 3.3)

Large-scale programmes or high-stakes decisions: Commission Advanced Methods (Chapter 3.4) using specialist support

Technical Annexes provide implementation templates for Foundation and Intermediate Methods, and commissioning specifications for Advanced Methods. Use Annexes as reference tools, not sequential reading.

A Glossary of Key Evaluation Terms is also provided at the beginning of the Annex for readers unfamiliar with technical evaluation terminology.

Case Studies (Section II) demonstrate methods in practice across diverse social service contexts.

Chapter 2:

Data governance and management standards



CHAPTER 2: DATA GOVERNANCE AND MANAGEMENT STANDARDS

2.1 Data as the Foundation for Evaluation

Data systems provide the evidentiary foundation for all evaluation activities. Without systematic data collection, programmes cannot demonstrate outcomes, assess effectiveness, or meet accountability requirements. Data governance establishes standards ensuring data collection is proportionate, legally compliant, ethically sound, and analytically useful.

This chapter establishes principles for data planning, collection, quality management, and legal compliance. Detailed protocols, templates, and implementation guidance appear in Annex A.

The proportionality principle applies to data governance: Data collection effort must match programme scale and evaluation requirements. Small programmes require basic systems addressing essential questions; large programmes require comprehensive systems supporting advanced evaluation methods. Over-collection burdens staff and participants without analytical benefit; under-collection prevents programmes demonstrating value. Data systems should collect minimum information necessary to answer evaluation questions whilst meeting funder and regulatory requirements.

2.2 Minimum Viable Data Requirements

All social service programmes regardless of scale must collect data addressing four fundamental questions:

Who participated? Basic demographic data enabling description of population served (target groups) and assessment of whether programme reaches intended beneficiaries. Typical variables: age, gender, referral source, baseline need or eligibility criteria. Data minimisation principle applies—collect only demographics necessary for evaluation or reporting requirements.

What did they receive? Participation data documenting service intensity and completion. Typical variables: enrolment date, sessions attended, programme completion status, dropout reasons where applicable. Enables assessment of engagement patterns and dose-response relationships.

Did outcomes improve? Outcome data measured at baseline and follow-up using validated instruments where available or fit-for-purpose indicators where validated measures do not exist. Outcome selection must align with Theory of Change (Chapter

3.2.1). Programmes should focus on 2-3 primary outcomes rather than attempting comprehensive measurement of all possible effects.

What did participants and stakeholders think? Feedback data capturing participant satisfaction, perceived benefit, implementation barriers, and suggestions for improvement. Stakeholder Feedback (Chapter 3.2.3) provides essential implementation intelligence complementing quantitative outcome data. Brief structured feedback (5-10 questions) typically sufficient for Foundation Methods; detailed qualitative research required only for Advanced Methods.

These four questions constitute Minimum Viable Data—the essential information every programme must collect to demonstrate basic accountability and enable learning. Programmes implementing only Foundation Methods need collect no additional data beyond these four categories unless funder requirements specify otherwise.

2.3 Data Planning Standards

Data collection must be planned systematically, not implemented opportunistically. Effective data planning requires:

Theory of Change as planning foundation: Data requirements derive directly from Theory of Change (Chapter 3.2.1). Activities specified in Theory of Change determine participation variables to track. Outcomes identified in Theory of Change determine outcome measures needed. Assumptions stated in Theory of Change determine feedback questions to ask stakeholders. Theory of Change prevents unfocused data collection addressing questions unrelated to programme logic.

Proportionate indicator selection: Theory of Change may identify numerous potential outcomes but programmes cannot measure everything. Indicator selection must be proportionate to programme scale and evaluation method employed. Foundation Methods (Chapter 3.2) require 2-3 primary outcome indicators. Intermediate Methods (Chapter 3.3) may add 3-5 secondary indicators. Advanced Methods (Chapter 3.4) require comprehensive indicator frameworks but only when specialist evaluation support enables sophisticated analysis justifying additional data burden.

Validated measures prioritised: Where validated outcome measures exist with established reliability and validity evidence, these should be employed. Validated measures enable comparison across programmes and strengthen credibility of findings. Where validated measures do not exist for relevant outcomes, fit-for-purpose indicators may be developed but should meet basic measurement standards: clear operational definitions, feasible data collection procedures, adequate sensitivity to detect change.

Integration into workflow: Data collection integrated into routine service delivery imposes minimal additional burden on staff and participants. Intake assessments

collect baseline demographic and outcome data. Service delivery documentation tracks participation. Exit procedures collect follow-up outcome data. Integration requires modest upfront design effort but dramatically improves data completeness and quality compared to retrospective data collection.

Realistic follow-up protocols: Outcome data requires follow-up measurement but programmes frequently lose contact with participants. Follow-up protocols must be realistic: collect multiple contact methods at intake (phone, email, emergency contact), schedule follow-up at reasonable intervals (3-6-12 months standard), offer modest incentives where appropriate, maintain equal effort for all participants to prevent attrition bias. Target follow-up rates of 70% or higher; rates below 50% substantially compromise validity and may require sensitivity analysis for attrition bias.

2.4 Data Quality Management Standards

Data quality determines evaluation credibility. Poor-quality data produces misleading conclusions regardless of analytical sophistication. Quality management requires systematic attention to completeness, consistency, accuracy, and timeliness.

Completeness: Completeness: Data collection procedures should achieve >70% completion rates for critical variables (consistent with the follow-up rate targets in Section 2.3). Missing data introduces bias when missingness is non-random (participants with poor outcomes are more likely to be lost to follow-up). Completeness monitoring should occur monthly, not only at evaluation end when attrition cannot be corrected. Programmes with persistent low completion rates should investigate causes (burden excessive, timing inappropriate, incentives inadequate) and adjust procedures.

Consistency: Staff must collect data identically using standardised protocols and instruments. Inconsistent data collection (different questions asked, different recording procedures, subjective interpretation) produces unreliable data unsuitable for analysis. Consistency requires: written data collection protocols, staff training on procedures, standardised data collection instruments (forms, surveys), periodic quality checks comparing staff data entry.

Accuracy: Data recorded should reflect reality. Common accuracy problems include: transposition errors in manual data entry, copy-paste errors duplicating records, coding errors assigning wrong categories, implausible values not flagged during entry (age 120 years, income €1 million monthly). Accuracy checking requires: validation rules in data systems preventing impossible values, spot-checking random sample of records against source documents, flagging outliers for verification.

Timeliness: Data should be entered promptly (within one week of collection) whilst events remain fresh in staff memory. Delayed data entry produces errors as staff forget

details or lose source documents. Timeliness also enables real-time quality monitoring identifying problems whilst corrective action remains possible.

Quality assurance processes: Systematic quality checking should occur monthly using standard checklists. Quality checks should examine: completion rates by variable and time period, consistency checks identifying anomalies, accuracy spot-checks on random sample, timeliness review of data entry lag. Quality issues identified should trigger corrective action (additional staff training, protocol revision, data re-checking) not merely documentation.

PRACTITIONER ALERT

When 70% Rate Is Difficult to Achieve

Some programme contexts face structural barriers to achieving 70% baseline-to-follow-up completion rates. Homelessness services, refugee and asylum seeker programmes, crisis intervention, outreach to street-connected populations, and programmes serving people with chaotic life circumstances routinely experience high participant mobility and disengagement. social service contexts.

This does not mean evaluation is impossible — it means design must be adapted:

Frontload data collection: Capture baseline data at first meaningful contact, not at a later "intake session" that some participants will never reach. Even a 5-minute baseline administered at first engagement dramatically improves completion rates.

Minimise respondent burden: Use the shortest validated instruments available (e.g., PHQ-2 rather than PHQ-9 for depression screening; single-item wellbeing measures where multi-item scales are impractical). Five minutes of data is infinitely more useful than a 30-minute questionnaire that most participants abandon.

Use administrative data linkage: Where participants consent and data-sharing agreements permit, link to administrative records (employment records, health service use, housing status) that do not depend on direct follow-up contact. This is particularly valuable for long-term outcome tracking.

Build data collection into service delivery: Integrate outcome measurement into routine service interactions (regular check-ins, review meetings, case management sessions) rather than treating evaluation as a separate activity requiring additional participant time.



Report completion rates transparently: When completion rates fall below 70%, evaluation does not become invalid but its conclusions must be qualified. Report achieved completion rates, analyse whether completers differ systematically from non-completers on available baseline variables, and discuss the direction of potential bias (e.g., "if those who dropped out had worse outcomes, our results overstate programme effectiveness by approximately X%").

Completion rates below 70% do not invalidate evaluation. They require honest reporting of limitations and considered interpretation of findings. A programme serving chaotic populations that achieves 50% completion with transparent limitation analysis produces more credible evidence than one claiming 90% completion through selective counting.

2.5 GDPR and Data Protection Standards

All evaluation activities must comply with General Data Protection Regulation (GDPR) (European Union, 2016) and national data protection legislation. GDPR compliance rests on four foundational principles:

Lawful basis for processing: Evaluation activities require lawful basis under GDPR Article 6. Common bases for evaluation include: Consent (Article 6(1)(a)) for sensitive data, participant interviews, optional data collection—requires explicit, informed, freely given consent documented systematically; Legitimate Interest (Article 6(1)(f)) for outcome monitoring using programme data for service improvement—requires documented balancing test demonstrating legitimate interest outweighs data subject rights; Public Task (Article 6(1)(e)) for evaluation mandated by public funders or regulatory requirements—requires clear legal or regulatory mandate. Special category data (Article 9) covering health, ethnicity, religion, sexual orientation requires explicit consent or specific exemptions (scientific research, public health monitoring). Programmes should minimise special category data collection to essential variables only.

Data minimisation: Programmes must collect only data necessary for specified evaluation purposes. Data minimisation prevents collection of "nice to have" demographic data without clear evaluation use. Every data point collected should answer specific evaluation questions or meet explicit regulatory requirements. Data minimisation assessment should occur during planning: for each proposed variable, justify evaluation necessity; delete variables lacking clear purpose. Data minimisation also requires regular review: data initially necessary may become unnecessary as evaluation progresses; delete unnecessary data promptly.

Data security: Programmes have legal and ethical obligations protecting participant data. Security requirements include: password-protected digital files, locked physical storage for paper records, access controls limiting data access to authorised staff only, encryption for sensitive data, secure disposal when retention period expires. Security

measures must be proportionate to data sensitivity: anonymised aggregate data requires basic security; identifiable health data requires heightened security measures.

Participant rights: Participants have rights under GDPR that programmes must respect: right to be informed (transparent privacy notice explaining data use), right of access (participants can request copy of their data), right to rectification (correct inaccurate data), right to erasure ("right to be forgotten"—with limitations for research under Article 89), right to restrict processing, right to object to processing, right to data portability. Programmes should establish procedures responding to participant rights requests within GDPR-specified timeframes (typically one month).

Cross-border data transfers: Multi-country programmes transferring personal data outside the European Economic Area (EEA) require appropriate safeguards. EU-to-EU transfers follow standard GDPR procedures. EU-to-UK transfers permitted under adequacy decision. EU-to-third-countries require Standard Contractual Clauses (SCCs) or equivalent safeguards unless the destination country has an EU-approved adequacy decision. Even for EEA countries (Norway, Iceland, Liechtenstein) and Switzerland (which have separate bilateral agreements), verify compliance requirements. Programmes planning cross-border transfers must establish transfer mechanisms before data collection begins.

PRACTITIONER ALERT

GDPR Applies to Your Programme – Regardless of Size

Every programme collecting personal data from participants — including names, contact details, demographic information, outcome scores, and feedback — is processing personal data under GDPR. This applies whether you have 20 participants or 20,000, whether your data is in a sophisticated database or a spreadsheet on someone's laptop, and whether you consider your work "research" or "routine monitoring."

At minimum, you must: (a) have a lawful basis for processing data (typically consent or legitimate interest — see Section 2.5); (b) inform participants clearly about what data you collect, why, and how it will be used (a one-page privacy notice is sufficient for most programmes); (c) store data securely (password-protected files, locked physical storage); (d) collect only data you actually need for evaluation or reporting (data minimisation); and (e) delete data when you no longer need it.

Non-compliance is not merely a legal risk — it is an ethical failure.

2.6 Data Management Planning

Systematic data management requires documented plans addressing collection, storage, security, sharing, and archiving. Data Management Plans (DMPs) are mandatory for HORIZON Europe projects and recommended best practice for all programmes employing Advanced Methods.

DMP essential elements: DMPs should specify: data to be collected (variables, instruments, frequency), data storage location and security measures, access controls and authorised users, data sharing arrangements (with whom, under what conditions, what safeguards), retention period and archiving procedures, GDPR compliance measures. DMPs should be living documents updated at key milestones when collection or sharing arrangements change.

FAIR principles for research data: Where evaluation produces data potentially useful beyond immediate programme, FAIR principles should guide data management: Findable (persistent identifiers, metadata), Accessible (clear access procedures, sustainable repositories), Interoperable (standard formats, common vocabularies), Reusable (clear licences, comprehensive documentation) (Wilkinson et al., 2016). Where feasible, anonymised datasets should be made openly available to support transparency, reproducibility, and secondary research, consistent with GDPR requirements and participant consent conditions. FAIR principles particularly relevant for multi-site evaluations, publicly-funded research, or evaluations producing generalisable knowledge beyond immediate programme context.

Data retention and disposal: Programmes should establish retention schedules specifying how long data will be kept. Retention requirements vary by purpose: funder requirements typically specify minimum retention (often 5-7 years post-programme completion), research ethics may require longer retention enabling verification, GDPR requires deletion when no longer necessary for original purpose. Identifiable data should be anonymised or deleted once no longer needed; aggregate data may be retained indefinitely.

2.7 Implementation and Resources

Data governance principles apply universally but implementation varies by programme scale. Small programmes implement basic data governance using simple tools (spreadsheets, basic survey platforms) and internal capacity. Medium-large programmes require structured data systems, dedicated data management staff or functions, and potentially specialist data protection advice.

For small programmes: Minimum viable data collected using simple forms integrated into intake/exit procedures. Basic spreadsheet sufficient for data management. Simple consent forms adapted from templates (Annex A.2). Monthly quality checks using checklists (Annex A.3). GDPR compliance through consent and basic security measures.

For medium programmes: Systematic data collection using standardised instruments. Basic database or survey platforms. Data Management Plan documented (Annex A.1). Regular quality monitoring. Consultation with data protection officer or legal counsel on GDPR compliance. Staff training on data collection and security procedures.

For large programmes and Advanced Methods: Comprehensive data systems supporting complex evaluation designs. Dedicated data management staff or functions. Sophisticated Data Management Plans addressing FAIR principles. Independent data quality audits. Legal review of data protection compliance. Formal research ethics review where required. Data sharing agreements for multi-site programmes.

Resources

- **Annex A.1:** Data Management Plan template
- **Annex A.2:** GDPR-compliant consent forms
- **Annex A.3:** Data quality checklists
- **Annex A.4:** Cross-border data transfer protocols

Chapter 3:

Methodological standards and selection



CHAPTER 3: METHODOLOGICAL STANDARDS AND SELECTION

3.1 Method Selection Framework

This chapter presents evaluation methods organised into three tiers reflecting programme scale, decision stakes, and analytical requirements. All programmes regardless of scale must implement Foundation Methods (Section 3.2). Established programmes add Intermediate Methods (Section 3.3) where efficiency questions or resource allocation decisions justify evaluation investment. Large-scale programmes or high-stakes policy decisions require Advanced Methods (Section 3.4) providing definitive causal evidence or comprehensive economic valuation.

Proportionality drives method selection: evaluation effort and sophistication must match programme scale, decision stakes, and available resources. Under-evaluation of major programmes fails accountability standards; over-evaluation of small programmes misallocates resources without commensurate analytical benefit.

3.1.1 How to Use This Chapter

Method selection follows three steps:

Step 1: Use this selection framework (Section 3.1) to identify appropriate methods based on programme scale, primary evaluation questions, and data availability.

Step 2: Read relevant method sections (3.2-3.5) for principles and standards. Each method section provides: definition and appropriate use cases, key principles, quality standards, common errors, implementation requirements. Main text provides standards and principles; detailed protocols appear in Annexes B.1-B.8.

Step 3: Consult Case Studies (Section II) for method demonstrations showing Foundation, Intermediate, and Advanced Methods implemented across diverse social service contexts.

Step 4: Use Annexes for implementation or commissioning. Foundation Methods (Annexes B.1-B.3) provide complete implementation templates. Intermediate Methods (Annexes B.4-B.6) provide partial guidance plus commissioning specifications. Advanced Methods (Annexes B.7-B.8) provide commissioning specifications for external specialists, not DIY implementation

This chapter need not be read sequentially. Navigate directly to sections addressing your evaluation needs.

3.1.2 Three-Tier Framework Overview

TIER 1: FOUNDATION METHODS (Section 3.2)

Mandatory for all programmes regardless of scale

Theory of Change (3.2.1): Maps how programme activities produce intended outcomes through specified causal pathways. Identifies assumptions, critical success factors, and external dependencies. Required for all programmes as foundation for all subsequent evaluation.

Outcome Monitoring (3.2.2): Systematic tracking of participant outcomes using validated measures where available or fit-for-purpose indicators where validated instruments do not exist. Documents whether outcomes improve from baseline to follow-up.

Stakeholder Feedback (3.2.3): Structured collection and analysis of participant voice, staff perspectives, and partner feedback. Provides implementation intelligence complementing quantitative outcome data.

TIER 2: INTERMEDIATE METHODS (Section 3.3)

For established programmes with efficiency questions or resource allocation decisions

Cost-Effectiveness Analysis (3.3.1): Compares costs per unit outcome across alternative delivery approaches. Answers "which approach achieves outcome X most efficiently?" Requires external evaluation support (health economists, evaluation specialists).

Multi-Criteria Decision Analysis (3.3.2): Structures complex decisions involving trade-offs between competing objectives, stakeholder values, and uncertain outcomes. Makes decision logic transparent and defensible. Requires facilitation expertise.

Social Return on Investment (3.3.3): Stakeholder-driven framework measuring social, environmental, and economic value creation. Extends cost-effectiveness principles with participatory outcome identification and valuation. Common in social investment contexts (impact bonds, social finance). Requires accredited practitioners.

TIER 3: ADVANCED METHODS (Section 3.4)

For large-scale programmes or high-stakes decisions requiring definitive evidence

Cost-Benefit Analysis (3.4.1): Comprehensive economic evaluation quantifying all costs and benefits in monetary terms. Gold standard for major policy decisions per UK Treasury Green Book and EU Better Regulation Guidelines. See Chapter 1.5 for CBA's unique positioning—comprehensive but resource-intensive, reserved for contexts where breadth justifies substantial investment in health economists or public policy analysts.

Quasi-Experimental Design (3.4.2): Uses comparison groups and statistical methods to establish whether programmes cause observed outcomes. Employs propensity score matching, difference-in-differences, regression discontinuity, or instrumental variables controlling for selection bias. Requires causal inference specialists.

Randomised Controlled Trials (3.4.3): Random assignment creates equivalent treatment and control groups, eliminating selection bias. Gold standard for causal inference. Requires trial methodologists, statisticians, research infrastructure. Substantial resource requirements and ethical considerations.

Realist Evaluation (3.4.4): Understands how programmes work in complex contexts by identifying context-mechanism-outcome configurations. Incorporates Contribution Analysis for complex multi-actor environments. Requires skilled qualitative researchers with complexity expertise.

3.1.3 Method Selection by Primary Evaluation Question

Match methods to questions you need answered:

Table 6 - Method Selection by Evaluation Question

Evaluation Question	Method	Section
How does our programme create change?	Theory of Change	3.2.1
Are outcomes improving for participants?	Outcome Monitoring	3.2.2
What do participants and stakeholders think?	Stakeholder Feedback	3.2.3
Which delivery approach achieves outcomes most efficiently?	Cost-Effectiveness Analysis	3.3.1
How should we choose between options with multiple competing criteria?	Multi-Criteria Decision Analysis	3.3.2
What social value per unit invested do stakeholders perceive?	Social Return on Investment	3.3.3
Does this programme create more value than it costs to society?	Cost-Benefit Analysis	3.4.1
Did this programme cause observed outcomes?	Quasi-Experimental Design or RCT	3.4.2, 3.4.3
How, why, for whom, and in what contexts does this work?	Realist Evaluation	3.4.4

3.1.4 Method Selection by Programme Scale

Small-scale programmes:

- **Mandatory:** Foundation Methods (Theory of Change, Outcome Monitoring, Stakeholder Feedback)
- **Consider adding:** Basic Cost-Effectiveness Analysis if comparing delivery models internally
- **Not proportionate:** Advanced Methods requiring substantial specialist investment

Medium-scale established programmes:

- **Mandatory:** Foundation Methods
- **Add where relevant:** Cost-Effectiveness Analysis (comparing delivery approaches), Multi-Criteria Decision Analysis (resource allocation decisions), SROI (when funders require social value demonstration or for social investment contexts)
- **Consider commissioning:** Advanced Methods only if decision stakes justify investment or funder requires

Large-scale programmes or high-stakes decisions:

- **Mandatory:** Foundation Methods
- **Typically required:** At least one Intermediate Method
- **Commission as appropriate:** Advanced Methods providing definitive causal evidence (QED, RCT) or comprehensive economic valuation (CBA) when decision stakes justify specialist investment

3.1.5 Data Requirements by Method

Select methods matching data collection capacity:

Table 7 - Data Requirements by Method

Method	Quantitative Data	Qualitative Data	Comparison Group	Typical Duration
Theory of Change	Optional	Required	No	1-2 weeks
Outcome Monitoring	Required	Optional	No	Ongoing
Stakeholder Feedback	Optional	Required	No	Quarterly
Cost-Effectiveness Analysis	Required	Optional	No	1-2 months

Multi-Criteria Decision Analysis	Mixed	Mixed	No	3-5 days
Social Return on Investment	Required	Required	No	1-3 months
Cost-Benefit Analysis	Required	Optional	No	2-4 months
Quasi-Experimental Design	Required	Optional	Required	6-12 months
Randomised Controlled Trial	Required	Optional	Required	18-36 months
Realist Evaluation	Optional	Required	No	12-24 months

3.1.6 Combining Methods

Methods may be combined when single approaches cannot answer all evaluation questions:

Foundation + Intermediate: Theory of Change + Outcome Monitoring + Cost-Effectiveness Analysis enables comparison of delivery approaches with documented outcomes. Example: youth employment programme tracking outcomes whilst comparing costs of intensive versus standard support.

Foundation + Intermediate (Social Value): Theory of Change + Outcome Monitoring + Social Return on Investment demonstrates programme logic with stakeholder-valued outcomes. Example: Social enterprise tracking participant outcomes whilst calculating social value for impact investors.

Causal + Economic: Randomised Controlled Trial + Cost-Benefit Analysis proves programme causes outcomes whilst calculating economic returns. Example: Housing First programme using randomisation plus comprehensive benefit monetisation.

Causal + Process: Quasi-Experimental Design + Realist Evaluation establishes whether a programme works whilst understanding how, why, and in what contexts. Example: mental health intervention with comparison group plus mechanism exploration.

Quantitative + Qualitative: Outcome monitoring + case studies provides outcome patterns plus rich understanding of participant experiences. Example: cross-border programme with standardised outcomes plus country-specific qualitative research.

Method combinations require additional resources and expertise. Ensure combined approach serves clear analytical purpose, not merely comprehensive documentation.

3.1.7 Quality Principles Across All Methods

Regardless of tier or method, high-quality evaluation requires:

Proportionality: Match sophistication to programme scale and decision stakes. Over-evaluation wastes resources; under-evaluation fails accountability standards.

Transparency: Document assumptions, data sources, analytical choices, and limitations explicitly. Enable replication by independent analysts.

Stakeholder engagement: Involve participants, families, communities, and delivery partners meaningfully in evaluation design, data interpretation, and use of findings.

Technical rigour: Implement methods according to established standards. Statistical analyses follow current practice. Sample sizes adequate for intended inferences. Causal claims justified by appropriate designs.

Ethical conduct: Protect vulnerable populations whilst respecting autonomy and dignity. Obtain informed consent. Manage data securely. Evaluation should not impose excessive burden on participants.

Fitness for purpose: Choose methods for answering priority questions with appropriate level of certainty. Methods should inform actual decisions, not merely document activities.

Independence: Evaluators should be independent from programme management to enable objective assessment. For small programmes where external evaluation is infeasible, internal evaluators should report to governance boards and findings should be subject to external peer review.

These principles are detailed in Chapter 1.7 and apply across all three tiers.

3.1.8 Reading This Chapter: Method Section Structure

Method sections (3.2-3.5) follow consistent structure:

Each method section contains:

1. Definition and appropriate use cases (when method applies, when unsuitable)
2. Key principles essential to the method
3. Quality standards and common errors
4. Implementation requirements (expertise needed, typical resources, timeline)
5. Resources (relevant Annexes, case studies, technical references)

Foundation Methods (3.2) provide complete main text guidance plus full implementation protocols in Annexes. Intermediate Methods (3.3) provide principles plus partial templates. Advanced Methods (3.4) provide standards for commissioning external specialists, not DIY implementation guidance—detailed protocols in Annexes serve as commissioning specifications, not practitioner manuals.

3.2 FOUNDATION METHODS (Mandatory Standards)

Foundation Methods are mandatory for all social service programmes regardless of scale. These methods provide essential building blocks for learning and accountability: Theory of Change establishes causal logic, Outcome Monitoring tracks whether change occurs, and Stakeholder Feedback captures participant voice. Every programme must implement these methods proportionate to scale and complexity.

Foundation Methods can be implemented using internal capacity with modest external support for method design where needed. Complete implementation protocols, templates, and quality assurance frameworks appear in Annexes B.1-B.3.

3.2.1 Theory of Change

Definition and Purpose

Theory of Change explicates the causal assumptions underlying social service interventions. Every programme operates on implicit theories about how activities generate outcomes; Theory of Change makes these theories visible, testable, and improvable. Theory of Change maps causal pathways linking programme inputs through activities and outputs to outcomes and impacts, whilst making explicit the assumptions underpinning each causal link.

Theory of Change differs from logic models in depth and testability (Weiss, 1995). Logic models provide linear representation (inputs → activities → outputs → outcomes → impacts). Theory of Change adds causal mechanisms, contextual factors, alternative pathways, and explicit assumptions enabling testing whether theoretical propositions hold true in practice.

When Theory of Change is Required

Theory of Change is mandatory for all programmes as foundation for all subsequent evaluation. Theory of Change determines what to measure (outcomes identified in causal pathways), when to measure (outcome timeframes specified), what data to collect (variables tracking assumptions), and how to interpret findings (outcomes contextualised within causal logic).

Theory of Change is particularly essential for: planning new programmes or major redesigns, designing evaluation strategies and identifying measurement priorities,

complex multi-component interventions with multiple outcome pathways, building stakeholder consensus about programme strategy, innovation contexts requiring learning about what works and why, EU funding applications requiring clear results frameworks.

Even simple programmes benefit from Theory of Change development. Investment of 1-2 days prevents months of confusion about measurement and interpretation.

Key Principles

Participatory development: Theory of Change must involve multiple stakeholders—front-line staff, participants or representatives, management, partners, governance representatives. Front-line staff possess practice wisdom often missing from management perspectives. Service users offer insights about what actually matters for outcomes. Participatory development builds shared understanding and ownership whilst incorporating diverse perspectives on causal pathways and assumptions.

Evidence-based causal logic: Theory of Change should ground causal pathways in research evidence about what works, practice wisdom from experienced practitioners, and lived experience insights from service users. Causal links must be plausible—supported by evidence or clear logic. Balance ambition with realism about achievable outcomes given programme resources and duration.

Explicit and testable assumptions: Assumptions must be stated clearly and specifically, not vaguely. Vague statements like "training improves employment" provide no testable propositions. Explicit assumptions specify conditions required for causation: "Participants completing 80% or more of job skills training sessions will demonstrate measurable improvements in interview performance and application quality (assessed by independent evaluators), leading to increased job offers within three months." Explicit assumptions enable testing whether theoretical propositions hold true.

Assumption prioritisation: Not all assumptions are equally important. Critical assumptions are those most essential to programme success AND most uncertain. High-priority assumptions require testing through data collection and monitoring. Low-priority assumptions (well-evidenced or less critical to success) require less systematic testing. Assumption prioritization focuses limited evaluation resources on most important questions.

Proportionate complexity: Theory of Change sophistication must match programme scale and complexity. Small programmes require simple one-page visual models with 3-5 key causal pathways and 5-10 critical assumptions. Large multi-component programmes need more detailed mapping with multiple pathways, evidence review, written narrative (5-10 pages), and comprehensive assumption testing matrix. Focus on



most important causal pathways rather than exhaustive detail capturing every possible connection.

Contextual factors explicit: Programmes operate within contexts affecting whether interventions work—economic conditions (labour market opportunities, benefit system rules), policy environment (regulatory requirements, complementary services), social and cultural factors (stigma, community norms), geographic factors (urban/rural setting, transport access), temporal factors (seasonal effects, current events). Theory of Change must map contextual enablers and constraints, not assume universal effectiveness regardless of context.

Living document: Theory of Change must be reviewed and updated at least annually based on emerging evidence from programme's own monitoring, new research on similar programmes, stakeholder feedback, implementation learning, and contextual changes. Theory of Change updated after major external shocks (policy changes, economic disruption, unexpected evaluation findings) even if annual review is not yet due. Theory of Change should be used actively in planning, evaluation design, staff training, and reporting—not produced once for funding application then filed away.

Core Elements

Inputs: Resources invested—funding, staff time, facilities, knowledge, partnerships, access to populations.

Activities: What programme actually does—training sessions, counselling, advocacy, service coordination, outreach, assessment, referral.

Outputs: Direct countable products of activities—participants served, sessions delivered, materials produced, referrals made, geographic coverage achieved. Outputs are not outcomes. Common error: confusing "120 participants complete training" (output) with "120 participants gain employment" (outcome).

Outcomes: Changes resulting from programme activities, distinguished by timeframe:

- Short-term outcomes (0-6 months): Knowledge gained, attitudes shifted, skills developed, immediate behaviour changes
- Medium-term outcomes (6-24 months): Sustained behaviour changes, new practices adopted, conditions improved, secondary effects emerging
- Long-term outcomes (2-5+ years): Sustained improvements, stable changed conditions, broader life changes

Impacts: Broader societal changes to which programme contributes alongside other factors—population wellbeing, social inclusion, economic security, system changes, policy reform. Programmes rarely cause impacts alone but contribute as one factor

among many. Attribution to single programmes typically inappropriate at impact level; contribution language more accurate.

Assumptions: Beliefs underlying each causal link that must hold true for causation to flow. Example: "Job skills training leads to employment" assumes (a) training content matches current employer needs, (b) participants attend and engage with content, (c) skills transfer from classroom to actual workplace settings, (d) local labour market has suitable opportunities for target population, (e) employers willing to hire target population despite barriers. Each assumption is a testable proposition requiring evidence.

Causal mechanisms: How change occurs, not merely what changes. Mechanisms explain the process by which activities produce outcomes. Example: skills training produces employment through mechanisms of: increased human capital (participants possess capabilities employers value), improved self-efficacy (participants believe they can succeed), enhanced signalling (credentials demonstrate capability to employers), expanded networks (programme connections facilitate opportunities). Understanding mechanisms enables programme improvement when outcomes do not materialise— which mechanism failed?

PRACTITIONER ALERT BOX

Outputs are Not Outcomes

This is the most common error in social service evaluation. Outputs describe what the programme produced, i.e. sessions delivered, participant enrolled, materials distributed. Outcomes describe what changed for the people served, i.e. employment gained, health improved, housing stabilised.

Confusing outputs with outcomes is not simply a labelling issue, it fundamentally undermines accountability of the evaluation. Funders and the public want to know whether lives improved, not whether activities occurred.

Example: A programme that delivers 50 training sessions to 500 participants demonstrate activity not value –hence show programme outputs. Showing that 120 of those participants reached sustained employment demonstrate change –hence reflect programme outcome.

When developing your Theory of Change and outcome monitoring system, test every proposed outcome by asking: Does this describe something we did, or something that changed for the people served? If it describes what you did, it is an output. If it describes what changed, it is an outcome.

Note that this distinction is fundamental to everything else in this book. Getting it wrong cascades through ToC, outcome monitoring, CEA, SROI, and CBA.



Quality Standards

Theory of Change must meet minimum quality standards:

Causal plausibility: Each causal link supported by evidence (research findings, practice wisdom, lived experience) or clear logical reasoning. Unsupported causal chains undermine credibility.

Distinguishes outputs from outcomes: Outputs describe programme products (what staff produce). Outcomes describe participant changes (what happens for people served). Confusion between these undermines evaluation validity.

Realistic outcome timeframes: Short-term outcomes achievable within 0-6 months. Medium-term outcomes require 6-24 months. Long-term impacts emerge over 2-5+ years. Unrealistic timeframes (expecting employment stability after a two-week programme) reflect poor causal logic.

Explicit testable assumptions: Every critical causal link accompanied by explicit assumptions specifying what must be true for causation to occur. Assumptions stated specifically enough to enable testing through data collection.

Assumption prioritisation documented: Critical assumptions (high importance AND high uncertainty) distinguished from well-evidenced or less critical assumptions. Evaluation priorities flow from this prioritisation.

Contextual factors mapped: External conditions enabling or constraining programme effectiveness explicitly identified—economic, policy, social, cultural, geographic, temporal factors.

Appropriate complexity: Simple programmes have simple Theory of Change (one-page visual, 3-page narrative). Complex multi-component programmes have detailed Theory of Change (multi-pathway visual, 5-10 page narrative, comprehensive assumption testing matrix). Complexity proportionate to programme sophistication.

Visual clarity: Visual representation fits one page (A3 maximum), uses consistent conventions (shapes for different element types, clear arrows showing causal direction), readable by non-experts (minimum 10-point font, jargon minimised, legend provided), accessible (colour-blind friendly palettes).

Stakeholder participation documented: Front-line staff (minimum 50% of staff team) and service users or representatives (ideally 20-30% of participants) contributed meaningfully to Theory of Change development through workshops or structured consultations.

Active use: Theory of Change referenced regularly in programme planning, evaluation design, staff training, progress reporting. Theory of Change displayed in staff areas. New staff receive a Theory of Change explanation during induction. Theory of Change is operational DNA, not a compliance document.

Regular updating: Theory of Change reviewed at least annually incorporating: new evidence from programme's own monitoring, research on similar programmes, stakeholder feedback about accuracy, implementation learning (where practice diverges from theory), contextual changes. Theory of Change version control maintained (dates, change logs, archived previous versions).

Common Errors

Error 1: Confusing outputs with outcomes. "120 participants complete training" is output (programme product). "120 participants gain employment" is the outcome (participant change). This confusion undermines evaluation validity because outputs do not demonstrate value creation.

Error 2: Vague untestable assumptions. "Training works" provides no testable proposition. "Participants completing 80% or more sessions will demonstrate measurable skill improvements" enables testing. Vague assumptions cannot guide data collection or interpretation.

Error 3: Linear thinking ignoring context. Programmes work differently in different contexts. Theory of Change assuming universal effectiveness regardless of economic conditions, policy environment, cultural factors, or implementation quality reflects poor causal reasoning. Contextual factors must be explicit.

Error 4: Static compliance document. Theory of Change developed for funding application then never referenced again wastes effort and prevents organisational learning. Theory of Change must be a living document actively used and regularly updated.

Error 5: Excessive complexity. Twenty-page Theory of Change documents with 50 outcomes and 200 assumptions become unusable. Focus on most important causal pathways and critical assumptions. The proportionality principle applies to the Theory of Change itself.

Error 6: No assumption prioritisation. Treating all assumptions as equally important fails to focus evaluation on critical uncertainties. High-priority assumptions (critical to success AND highly uncertain) require systematic testing. Low-priority assumptions require minimal attention.

Implementation Requirements

Expertise needed: Theory of Change development requires facilitation skills but not specialist technical expertise. Programme staff can lead the process with external facilitation support where helpful for stakeholder workshops. Small programmes are implemented using internal capacity. Medium-large programmes may commission external facilitators with Theory of Change expertise.

Time investment: Simple programmes (small scale, straightforward causal logic) require half-day stakeholder workshop plus 1-2 days documentation—total 1-2 weeks. Medium programmes require a full-day stakeholder workshop, evidence review, written narrative development—total 2-4 weeks. Large complex programmes require multiple workshops, literature review, detailed assumption testing matrix—total 4-6 weeks.

Stakeholder engagement: Minimum requirements: 50% of front-line staff participate, 20-30% service user representation (through representatives where direct participation is infeasible), management and governance involvement, key partner participation where the programme relies on external coordination.

Resources:

- **Annex B.1.1:** Theory of Change narrative template
- **Annex B.1.2:** Assumption testing matrix
- **Annex B.1.3:** Stakeholder consultation protocol
- **Annex B.1.4:** Visual design guidance and conventions
- **Annex B.1.5:** Quality assurance protocol with annual review process
- **Annex B.1.6:** Service-specific Theory of Change examples (education, employment, health, housing, family support)
- **Case Study 1:** Finland Youth Guarantee Theory of Change development process

Integration with other methods: Theory of Change provides foundation for all subsequent evaluation. Outcome Monitoring (Section 3.2.2) tracks outcomes identified in Theory of Change. Stakeholder Feedback (Section 3.2.3) tests assumptions through participant and partner perspectives. Cost-Effectiveness Analysis (Section 3.3.1) compares costs per Theory of Change outcome. All Advanced Methods build on Theory of Change causal logic.

3.2.2 Outcome Monitoring

Definition and Purpose

Outcome monitoring involves systematic collection and analysis of data tracking changes experienced by service users during and after programme participation. Outcome monitoring answers whether participants experience improvements identified in Theory of Change as intended programme outcomes.

Outcome monitoring tracks whether expected changes occur but cannot prove causation. Impact evaluation (Section 3.4.2, 3.4.3) uses comparison groups demonstrating programme caused observed changes. Outcome monitoring provides valuable performance evidence informing management decisions and accountability requirements, and may indicate where more rigorous causal evaluation is warranted.

When Outcome Monitoring is Required

Outcome monitoring is mandatory for all programmes regardless of scale. Programmes must demonstrate whether intended beneficiaries experience intended changes. Outcome monitoring provides evidence for: accountability to funders and commissioners, data-driven service delivery improvements, identifying which participants benefit most or least, detecting problems or unintended consequences early, building evidence base before investing in more sophisticated evaluation methods.

Key Principles

Alignment with Theory of Change: Outcome monitoring must directly reflect outcomes identified in Theory of Change (Section 3.2.1). This ensures measuring what matters most whilst avoiding data collection overload. Prioritise outcomes where: change is expected within programme timeframe, measurement is feasible with available resources, results will inform practical decisions.

Proportionate measurement: Select measurement approaches balancing rigour with practicality. A perfectly valid measure that is too burdensome to collect reliably is less useful than a simpler measure consistently implemented. Collect fewer measures more consistently rather than comprehensive data sporadically. Integrate data collection into routine service delivery where possible (intake assessments, case review meetings, exit procedures).

Before-and-after comparison: Outcome monitoring requires measuring outcomes at multiple time points tracking change. Minimum requirement: baseline measurement when participants enter service and follow-up measurement at programme completion or appropriate intervals. Multiple follow-up points (3 months, 6 months, 12 months) strengthen evidence by showing trajectories rather than single snapshots. Baseline measurement must occur before or at service commencement; retrospective baseline collection produces unreliable data.

Validated measures prioritised: Where validated outcome measures exist with established reliability and validity evidence, programmes should employ these instruments. Validated measures enable comparison across programmes and strengthen credibility of findings. Examples include standardised wellbeing scales such as the Warwick-Edinburgh Mental Well-being Scale (Tennant et al., 2007), Patient Health

Questionnaire-9 for depression (Kroenke et al., 2001), and Generalized Anxiety Disorder-7 scale (Spitzer et al., 2006), as well as employment outcome measures, housing stability indicators, and validated social support scales. Where validated measures do not exist for relevant outcomes or are inappropriate for target population, fit-for-purpose indicators may be developed but should meet basic measurement standards: clear operational definitions, feasible data collection procedures, adequate sensitivity to detect change.

Systematic procedures: Reliable monitoring depends on consistent data collection. Requirements include: standardised instruments and procedures ensuring comparability across participants and time periods, clear protocols specifying who collects data when and how, staff training on data collection procedures and instrument administration, quality assurance processes identifying data problems (missing data, inconsistent recording, implausible values), secure storage protecting participant confidentiality whilst enabling analysis.

Regular analysis and use: Outcome monitoring only adds value if data is actually analysed and used for decisions. Establish regular reporting cycles (monthly for large programmes, quarterly for medium programmes, annually minimum for small programmes) to: review outcome trends, analyse disaggregated by participant characteristics identifying who benefits most, create feedback loops connecting findings to service delivery decisions, communicate progress to staff and stakeholders.

Quality Standards

Outcome monitoring must meet minimum quality standards:

Theory of Change alignment: Outcomes monitored directly correspond to outcomes identified in Theory of Change. Monitoring system tracks 2-3 primary outcomes minimum (small programmes) to 5-7 outcomes (large programmes). Monitoring focuses on outcomes most central to programme logic and most feasible to measure reliably.

Baseline data collection: Baseline outcomes measured before or at programme commencement for all participants. Retrospective baseline ("how were you before starting?") is unreliable and should be avoided. Baseline completion rate >80% minimum; <70% compromises validity.

Follow-up data collection: Follow-up outcomes measured at programme completion minimum. Programmes with longer-term outcome objectives require extended follow-up (3, 6, 12 months post-completion). Follow-up completion rate >70% minimum; <50% severely compromises validity and introduces attrition bias.

Validated instruments where available: Validated outcome measures with published psychometric properties used where available and appropriate for target population.

Where validated measures are unavailable, fit-for-purpose indicators have clear operational definitions and documented measurement procedures.

Consistent administration: Data collection procedures standardised. All staff administering instruments trained on proper procedures. Administration timing consistent (e.g., baseline always at first contact, not varying between participants).

Data quality management: Regular quality checks identify missing data, inconsistent recording, implausible values. Quality monitoring occurs monthly (large programmes) to quarterly (small programmes), not only at year-end when problems cannot be corrected.

Disaggregated analysis: Outcomes analysed by relevant participant characteristics (age, gender, baseline severity, service intensity) identifying differential effects. Large disparities in outcomes across subgroups investigated.

Appropriate interpretation: Monitoring results interpreted acknowledging limitations. Changes observed might result from programme intervention, natural improvement over time, other services, external life changes, or measurement effects. Interpretation acknowledges these limitations whilst valuing evidence generated. Causal claims avoided unless employing comparison group designs (Section 3.4.2, 3.4.3).

PRACTITIONER ALERT

Outcome Monitoring Cannot Prove Causation

When the participants' outcomes improve from baseline to follow-up, it is encouraging. However, it does not prove your programme caused the improvement. Outcomes might have improved because of your programme, but also because of natural recovery (people tend to seek help when things are worst, then improve naturally), other services participants received simultaneously, changes in personal circumstances, economic conditions improving etc.

This does not mean outcome monitoring is worthless. It provides valuable evidence that participants' lives improved during programme engagement. It cannot tell, however, it improved definitely because of the programme. This question of causation requires comparison group designs (See sections 3.4.2 – 3.4.3).

In practice: Report monitoring findings as “participants experienced X improvement” rather than “the programme achieved X improvement”. If you need to show whether the programme caused outcomes (i.e. for scaling decisions), you will need to add other methods from Intermediate or Advanced Methods.



Common Errors

Error 1: Missing baseline data. Collecting baseline after participants have begun services makes measuring change impossible. Baseline must be collected before or at first service contact.

Error 2: Poor follow-up rates. Outcome monitoring fails if unable to track participants through to follow-up. Attrition often biases results—participants with poor outcomes are more likely to be lost to follow-up. Solution: integrate follow-up data collection into final service contacts, collect multiple contact methods at baseline, allocate resources for participant tracking, offer modest incentives for follow-up completion.

Error 3: Measuring too much. Attempting comprehensive measurement often results in measuring nothing well due to staff burden and participant fatigue. Solution: start with a focused set of 3-5 core outcome measures aligned with Theory of Change primary outcomes. Expand gradually as capacity develops.

Error 4: Data collection without use. Data collection that does not inform decisions wastes resources and burdens staff and participants. Solution: establish clear expectations for how monitoring data will be used before implementing systems, create regular review processes where findings inform action.

Error 5: Treating monitoring as causal evaluation. Outcome monitoring cannot prove service caused observed changes. Claiming causal impact based solely on monitoring data lacks credibility and misrepresents limitations. Solution: be transparent about what monitoring can and cannot demonstrate. If proving causality matters for scaling or policy decisions, plan for impact evaluation using comparison groups (Section 3.4.2, 3.4.3).

Error 6: Ignoring negative or null findings. Monitoring may reveal no improvement or deterioration for some participants. Defensive dismissal of unwelcome findings prevents learning. Solution: investigate reasons for poor outcomes, examine whether specific subgroups fare worse, consider programme modifications, acknowledge openly when interventions not producing intended effects.

Implementation Requirements

Expertise needed: Outcome monitoring requires data collection skills and basic quantitative analysis capacity but not specialist technical expertise. Programme staff can implement training on: instrument administration, data quality protocols, basic descriptive statistics. Small programmes are implemented using internal capacity. Medium-large programmes may require a dedicated data coordinator. External support helpful for: selecting validated instruments, designing data systems, staff training, initial analysis.



Time investment: Initial setup requires 2-4 weeks (instrument selection, data system design, staff training, piloting procedures). Ongoing effort: 2-3 hours monthly per 100 participants for data quality monitoring, quarterly analysis, annual comprehensive reporting.

Data requirements: Minimum data collection: participant identifiers enabling tracking, baseline outcome measures, follow-up outcome measures at appropriate intervals, basic demographic and service participation data enabling disaggregated analysis. See Chapter 2 for data governance standards.

Resources:

- **Annex B.2.1:** Outcome measurement planning template
- **Annex B.2.2:** Sample baseline and follow-up forms
- **Annex B.2.3:** Analysis spreadsheet templates
- **Annex B.2.4:** Reporting templates for different audiences
- **Annex B.2.5:** Validated measurement tools by service type (wellbeing, employment, housing, family support, health)
- **Annex B.2.6:** Data quality checklist and monitoring protocols
- **Case Studies 1, 5:** Outcome monitoring demonstrations

Integration with other methods: Outcome monitoring data provides essential inputs for Cost-Effectiveness Analysis (Section 3.3.1), enables Theory of Change assumption testing, informs selection of participants for qualitative case studies, supplies outcome data for comparison group designs (Section 3.4.2, 3.4.3).

3.2.3 Stakeholder Feedback

Definition and Purpose

Stakeholder feedback involves systematically collecting and analysing views, experiences, and satisfaction levels of people who have stake in social service programmes. Stakeholder feedback encompasses both qualitative insights (what people think and why) and quantitative ratings (satisfaction levels, perceived service quality). Stakeholder feedback focuses on perceptions, experiences, and satisfaction rather than objective outcome measurement.

Stakeholder feedback differs from outcome monitoring. Stakeholder feedback measures satisfaction and experience (subjective perceptions). Outcome monitoring tracks actual changes in participants' lives (objective measures of wellbeing, employment, skills, etc.). Both are essential. High satisfaction without outcome improvements suggests pleasant but ineffective service. Outcome improvements with low satisfaction indicate effective intervention delivered in an inaccessible or unacceptable manner.

When Stakeholder Feedback is Required

Stakeholder feedback is mandatory for all programmes as Foundation Method. Stakeholder perspectives provide: evidence of service quality and accessibility, early warning of implementation problems, insights enabling rapid service improvements, accountability evidence for commissioners valuing participant voice, contextual understanding complementing quantitative data, verification whether Theory of Change assumptions about participant engagement and acceptability hold true.

Key Principles

Multiple perspectives: Different stakeholders offer distinct insights essential for comprehensive understanding. Consulting only service users misses staff implementation challenges, partner collaboration issues, and funder strategic concerns. Balanced feedback systems incorporate: primary stakeholders (service users/participants, front-line staff, referral partners), secondary stakeholders where relevant (funders, commissioners, community representatives, governance boards). Minimum requirement: service user and staff feedback. Comprehensive systems add partner and funder perspectives.

Safe and accessible mechanisms: Stakeholders must feel able to provide honest feedback without negative consequences. Requirements include: anonymity or confidentiality where appropriate (particularly for service user satisfaction surveys), multiple feedback channels accommodating different preferences (surveys, interviews, focus groups, suggestion boxes, online platforms), plain language avoiding jargon, accessible formats (translation, large print, easy read), clear communication that feedback will not affect service access or employment.

Systematic collection: Ad hoc feedback is valuable but insufficient for Foundation Method requirements. Systematic approaches include: regular collection cycles (quarterly minimum, more frequent for rapidly changing programmes), standardised core questions enabling comparison over time and across stakeholder groups, representative sampling or census approaches, documented collection processes, consistent recording systems.

Balance between structure and openness: Effective feedback systems combine: structured elements (standardised satisfaction scales, specific questions about service components, quantifiable rating scales) with open elements (free-text comment fields, open-ended questions inviting unexpected insights, narrative accounts of experiences, creative methods for populations uncomfortable with surveys).

Closing the feedback loop: Stakeholder feedback only builds trust when demonstrating responsiveness. "You said, we did" communication requires: timely analysis identifying actionable findings, visible service improvements addressing feedback themes,

communication to stakeholders explaining what changed and why, honest explanation when unable to implement suggestions (resource constraints, regulatory requirements), celebration of improvements made through stakeholder input. Feedback without visible response teaches stakeholders their input does not matter, reducing future participation.

Quality Standards

Stakeholder feedback must meet minimum quality standards:

Multiple stakeholder groups consulted: Minimum: service users and front-line staff. Comprehensive systems add partners, funders, governance representatives. Each group consulted using appropriate methods.

Appropriate methods for stakeholder groups: Methods match stakeholder characteristics and programme context. Service users: brief exit surveys (5-10 questions), annual satisfaction surveys, focus groups (2-4 per year for medium-large programmes). Staff: annual staff surveys, regular team discussions, exit interviews for departing staff. Partners: annual consultation, relationship surveys. Methods accommodate literacy levels, language preferences, cognitive abilities, time constraints.

Good response rates: Service user surveys achieve >60% response rate minimum. Staff surveys achieve >70% response rate. Low response rates risk non-response bias where dissatisfied stakeholders disproportionately respond or non-respond. Strategies improving response rates: explain why feedback matters, minimise burden (brief instruments), provide multiple response options, offer incentives where appropriate, integrate into existing touch-points.

Safe and confidential: Anonymity provided for service user satisfaction surveys. Staff feedback collected through channels independent of line management where possible. Protection from retaliation explicit.

Regular collection cycles: Service users: exit feedback for all programmes, annual satisfaction surveys for programmes with long-term engagement. Staff: annual surveys minimum, quarterly pulse checks for large programmes. Partners: annual consultation minimum. Ad hoc feedback mechanisms (suggestion boxes, online forms) available continuously.

Structured and open questions: Instruments include: satisfaction rating scales (overall satisfaction, specific components), quantitative ratings enabling benchmarking and trend analysis, open-ended questions inviting detailed feedback ("What could we improve?", "What works well?"), space for unexpected insights not anticipated in structured questions.

Systematic analysis: Feedback analysed at least quarterly (large programmes) to annually (small programmes). Analysis identifies: satisfaction levels and trends, themes in qualitative comments, differences across stakeholder groups or service components, priorities for improvement (high importance, low satisfaction), positive aspects to maintain (high satisfaction).

Results inform decisions: Feedback findings discussed in team meetings, governance meetings, strategic planning processes. Specific service improvements traceable to stakeholder feedback. Documentation showing feedback loop: feedback received → analysis conducted → decisions made → changes implemented → stakeholders informed. This iterative cycle of feedback, analysis, decision, and adaptation constitutes adaptive management — the systematic use of evaluation evidence to adjust programme delivery in response to emerging findings and changing contexts.

Appropriate interpretation: High satisfaction does not prove effectiveness. Low satisfaction does not prove ineffectiveness. Satisfaction must be triangulated with outcome monitoring data. Stakeholder feedback represents perceptions which may or may not align with objective evidence but shape engagement and reputation regardless.

Common Errors

Error 1: Only seeking positive feedback. Designing instruments or timing collection to maximise positive responses (e.g., only surveying most engaged participants, collecting feedback immediately after enjoyable activities) produces misleading evidence. Solution: systematic sampling, neutral question wording, multiple collection points, anonymous channels encouraging honest criticism.

Error 2: Collecting feedback without action. Filing feedback away without analysis or response teaches stakeholders their input does not matter. Solution: establish clear processes from collection through analysis to action, communicate visible improvements, close feedback loop with "you said, we did" communications.

Error 3: Unrepresentative samples. Only hearing from satisfied participants or vocal critics, missing middle ground or marginalised voices. Solution: multiple methods reaching different groups, active outreach to non-responders, targeted effort to under-represented populations, consider accessibility barriers preventing feedback.

Error 4: Defensive responses to criticism. Dismissing negative feedback or blaming stakeholders for misunderstanding undermines trust and prevents learning. Solution: cultivate learning culture valuing feedback as improvement opportunity, investigate criticisms thoroughly, remember perception shapes stakeholder behaviour even when perception differs from staff perspective, thank stakeholders for honest feedback including criticism.

Error 5: Confusing satisfaction with effectiveness. High satisfaction indicates good experience but not necessarily good outcomes. Solution: always interpret stakeholder feedback alongside outcome monitoring data, distinguish process satisfaction (how service delivered) from outcome satisfaction (whether needs met), acknowledge when satisfaction is high but outcomes disappointing or vice versa.

Implementation Requirements

Expertise needed: Stakeholder feedback requires facilitation and qualitative analysis skills but not specialist technical expertise. Programme staff can implement training on: instrument design, focus group facilitation, qualitative coding, communicating findings. Small programmes implemented using internal capacity. Medium-large programmes may require external facilitators for focus groups (avoiding power dynamics where staff facilitate service user feedback).

Time investment: Initial setup: 1-2 weeks developing instruments and procedures. Ongoing effort: integrated into service delivery (exit surveys), quarterly analysis for medium-large programmes (4-8 hours per quarter), annual comprehensive analysis for all programmes (1-2 days).

Stakeholder engagement: Service user feedback: census approach (all participants) for exit surveys, representative samples for annual satisfaction surveys. Staff feedback: census approach (all staff). Partner feedback: key partners consulted annually. Governance feedback: annual or bi-annual depending on meeting frequency.

Resources:

- **Annex B.3.1:** Stakeholder mapping template
- **Annex B.3.2:** Participant satisfaction survey template
- **Annex B.3.3:** Focus group facilitation protocol
- **Annex B.3.4:** Staff feedback survey template
- **Annex B.3.5:** Feedback analysis framework (quantitative ratings and qualitative themes)
- **Annex B.3.6:** "You said, we did" communication templates
- **Case Study 5:** Stakeholder feedback demonstration

Integration with other methods: Stakeholder feedback provides essential validation of Theory of Change assumptions about acceptability and engagement, contextualises outcome monitoring findings, identifies implementation problems requiring investigation, informs Cost-Effectiveness Analysis by revealing process differences between delivery models, supplies participant voice for Realist Evaluation exploring contextual factors.

3.3 INTERMEDIATE METHODS (Standards and Commissioning Triggers)

Intermediate Methods provide efficiency evidence or support complex resource allocation decisions for established programmes with demonstrated implementation capability. Cost-Effectiveness Analysis compares costs per unit outcome across alternative delivery approaches. Multi-Criteria Decision Analysis structures complex decisions involving trade-offs between competing objectives. These methods typically require external evaluation support from health economists or decision analysts but remain feasible for medium-scale programmes where efficiency questions or allocation decisions justify evaluation investment.

3.3.1 Cost-Effectiveness Analysis

Definition and Purpose

Cost-Effectiveness Analysis (CEA) compares costs and outcomes of different interventions or service delivery models identifying which achieves outcomes most efficiently (Drummond et al., 2015). CEA measures outcomes in natural units (jobs secured, people housed, symptom reduction, quality-adjusted life years) rather than monetary values, calculating cost per outcome achieved. This enables comparison between alternative approaches to achieving the same outcome whilst avoiding contentious monetary valuation of social outcomes.

CEA expresses results as cost-effectiveness ratios: cost per outcome unit (e.g., cost per person employed, cost per quality-adjusted life year gained). Incremental Cost-Effectiveness Ratio (ICER) shows additional cost required to achieve one additional outcome unit by choosing one alternative over another.

When Cost-Effectiveness Analysis is Appropriate

CEA is appropriate when: comparing different models for achieving same outcome (intensive versus standard support, in-house versus commissioned delivery, different intervention components or intensities), making resource allocation decisions between programmes with similar objectives, optimising service design by comparing delivery variations, demonstrating efficiency to funders or commissioners requiring value-for-money evidence, scaling decisions requiring evidence which approach delivers outcomes most efficiently at scale.

CEA is inappropriate when: programmes serve fundamentally different populations (comparing youth employment with elderly care meaningless), outcomes measured differently across alternatives (one measures employment rate, another measures employability skills—not comparable), only one delivery model exists (CEA requires comparing at least two alternatives), causal attribution essential (CEA shows

association between costs and outcomes but cannot prove causation without comparison group designs—see Sections 3.4.2, 3.4.3).

Key Principles

Genuine comparability required: Alternatives must target the same population with the same eligibility criteria, aim for same outcomes using common measurement approach, operate in comparable contexts (cannot fairly compare urban intensive service with rural basic service), and have sufficient sample sizes enabling meaningful comparison. Comparing incomparable programmes produces misleading results.

Common outcome measurement: All alternatives must measure outcomes identically using same instruments, same timing, same definitions. If Programme A measures "employment rate at 6 months" and Programme B measures "employability skills improvement," they cannot be meaningfully compared through CEA.

Comprehensive cost accounting: Include all relevant costs for fair comparison: direct costs (staff delivering intervention, materials, participant incentives, venue hire), indirect costs (management, administration, facilities, IT infrastructure proportionally allocated), opportunity costs (alternative uses of resources particularly relevant for comparing in-house versus commissioned services), capital costs (equipment, vehicles, building modifications annualised over useful life). Exclude sunk costs (already spent regardless of decision). Include only costs varying between alternatives.

Incremental analysis: CEA focuses on differences between alternatives. $ICER = (\text{Cost Alternative A} - \text{Cost Alternative B}) / (\text{Outcome Alternative A} - \text{Outcome Alternative B})$. This answers: "How much extra does it cost to achieve one additional outcome unit by choosing Alternative A over Alternative B?" Incremental analysis is more informative than simple cost-effectiveness ratios when alternatives produce different outcome levels.

Perspective specification: CEA results differ depending on whose costs and outcomes counted. Programme perspective counts only costs borne by the programme itself. Funder perspective includes all costs the funder pays. Public sector perspective includes all government costs across departments. Societal perspective includes all costs to society including participants, families, volunteers. Specify perspective explicitly and justify choice. Societal perspective is the most comprehensive but most complex. Perspective choice affects which costs included and may affect interpretation of results.

Appropriate time horizon: Analysis timeframe should capture all relevant costs incurred, meaningful outcome measurement period, and sustained effects where relevant. Typical time horizons: 1-2 years for short-term interventions, 3-5 years for medium-term programmes, 10+ years for life-changing interventions. Too-short time

horizons miss delayed costs or benefits. Too-long time horizons introduce excessive uncertainty. Justify the time horizon based on when outcomes meaningfully emerge.

Discounting future values: Costs and outcomes occurring in future should be discounted to present values reflecting time preference and opportunity cost of capital. EU standard: 3% social discount rate per EU Better Regulation Vademecum (European Commission, 2021). UK standard: 3.5% discount rate per UK Treasury Green Book (HM Treasury, 2022). Apply the same discount rate to costs and outcomes unless strong justification for differential discounting.

PRACTITIONER ALERT

Which Discount Rate to Use?

Discounting adjusts the value of future costs and benefits to today's terms. The logic is straightforward: €1,000 received today is worth more than €1,000 received in five years, because today's money can be invested, and because people generally prefer benefits sooner rather than later.

For EU-funded programmes, use the 3% social discount rate specified in the EU Better Regulation Vademecum. For programmes evaluated under UK frameworks, use the 3.5% rate per the UK Treasury Green Book. If your programme operates under both frameworks (e.g. a UK partner in a HORIZON Europe project), use 3% for EU reporting and note the UK standard in sensitivity analysis.

For financial analysis (e.g. assessing organisational budget impacts as distinct from social value), the EU Vademecum specifies 4% financial discount rate.

Uncertainty management: CEA involves numerous assumptions and estimates. Address uncertainty through: sensitivity analysis testing how results change when varying key assumptions, scenario analysis exploring optimistic/pessimistic parameter combinations, confidence intervals where data permit statistical analysis, threshold analysis identifying parameter values at which conclusions change.

Quality Standards

CEA must meet minimum quality standards:

Alternatives genuinely comparable: Serve same population, pursue same outcomes measured identically, operate in comparable contexts. Comparability justified explicitly.



Comprehensive cost data: All relevant cost categories identified systematically. Costs measured accurately from financial records and activity data. Overhead and indirect costs allocated proportionally. Excluded costs justified explicitly.

Reliable outcome data: Outcomes measured using validated instruments where available or well-justified fit-for-purpose indicators. Outcome measurement timing consistent across alternatives. Adequate sample sizes. Missing data addressed appropriately.

Clear perspective: Analysis perspective (programme, funder, public sector, societal) stated explicitly and applied consistently. Cost inclusions/exclusions justified by perspective choice.

Appropriate time horizon: Time horizon captures meaningful costs and outcomes. Justification provided for time horizon choice. If different time horizons are tested in sensitivity analysis, rationale provided.

Discounting applied correctly: Future costs and outcomes discounted to present value. Discount rate specified (3% EU standard or 3.5% UK standard) and applied consistently.

Sensitivity analysis conducted: Key assumptions varied systematically. Results reported showing robustness or sensitivity to assumptions. Threshold analysis identifies tipping points where conclusions change.

Transparent reporting: Methods documented enabling replication. Data sources specified. Assumptions stated explicitly. Limitations acknowledged. Results presented clearly with appropriate caveats.

Common Errors

Error 1: Comparing incomparable programmes. Programmes must target the same population and measure the same outcomes. Comparing youth employment with elderly care is not meaningful CEA. Solution: ensure genuine comparability or conduct separate evaluations.

Error 2: Incomplete cost accounting. Omitting indirect costs or overhead biases results, typically favouring smaller interventions with lower apparent costs. Solution: systematic identification of all cost categories, proportional allocation of shared costs, transparency about included/excluded costs with justification.

Error 3: Short time horizons missing important effects. Some interventions have upfront costs but long-term benefits (or vice versa). Too-short time horizons misrepresent cost-effectiveness. Solution: justify time horizon based on when outcomes



meaningfully emerge, conduct sensitivity analysis with longer time horizons, acknowledge if evaluation occurs before full benefits materialised.

Error 4: Treating CEA as causal evaluation. CEA shows association between costs and outcomes but cannot prove programme caused outcomes without comparison group controlling for selection effects (Sections 3.4.2, 3.4.3). Solution: be transparent that CEA compares alternatives but does not establish causation, consider adding comparison group design if causal claims are essential.

Error 5: Over-interpreting point estimates. Single ICER figures are misleadingly precise given uncertainty in cost and outcome estimates. Solution: always conduct sensitivity analysis, present results as ranges not point estimates, acknowledge uncertainty explicitly.

Implementation Requirements

Expertise needed: CEA requires health economics or evaluation economics expertise including: cost data collection and analysis, outcome measurement, incremental analysis, sensitivity analysis, economic evaluation reporting. Small-medium programmes typically commission external health economists or evaluation specialists. Large programmes may have internal analytical capacity but benefit from external quality assurance.

Time investment: 1-3 months typical depending on data availability and complexity. Data collection phase: 2-4 weeks identifying cost categories, extracting financial data, collecting outcome data. Analysis phase: 2-4 weeks calculating ratios, conducting sensitivity analysis. Reporting phase: 1-2 weeks preparing findings for decision-makers.

Data requirements: Cost data: financial records for all cost categories, activity data enabling cost allocation, participant numbers. Outcome data: common outcome measures across all alternatives, adequate sample sizes (minimum 50 participants per alternative for reliable comparison, larger samples if outcome variance high), timing matched across alternatives. See Chapter 2 for data governance standards.

Commissioning external specialists: When commissioning CEA, specify: alternatives to compare, perspective for analysis, time horizon, outcome measures, available cost data, timeline for analysis, deliverables expected. Provide evaluators with: programme descriptions for all alternatives, financial records or budgets, outcome data or data collection protocols, access to programme staff for clarification. Reference Annex B.4 technical specifications requiring compliance with analytical protocols and quality standards.

Resources:

- **Annex B.4.1:** CEA planning template
- **Annex B.4.2:** Cost and outcome data collection templates
- **Annex B.4.3:** Cost calculation spreadsheet with sensitivity analysis
- **Annex B.4.4:** Results presentation templates
- **Annex B.4.5:** Quality assurance protocol
- **Case Study 2:** Individual Placement and Support (IPS) CEA demonstration

Integration with other methods: CEA builds on Theory of Change identifying outcomes to measure and Outcome Monitoring providing outcome data. CEA complements Cost-Benefit Analysis (Section 3.4.1) when some outcomes resist monetary valuation. CEA informs Multi-Criteria Decision Analysis (Section 3.3.2) providing cost-effectiveness as one criterion among many. CEA complements experimental designs (3.4.2) and randomised controlled trials (3.4.3) to evaluate the economic effectiveness of the interventions.

3.3.2 Multi-Criteria Decision Analysis

Definition and Purpose

Multi-Criteria Decision Analysis (MCDA) provides a structured framework for evaluating options against multiple, often conflicting, criteria when no single metric can capture all relevant considerations (Belton & Stewart, 2002). MCDA makes explicit the values and trade-offs underlying programme decisions, enabling transparent defensible choices even when evidence is incomplete.

MCDA systematically evaluates alternatives against multiple criteria, assigns weights reflecting relative importance of each criterion, scores how well each alternative performs on each criterion, and calculates weighted scores identifying preferred option(s).

When Multi-Criteria Decision Analysis is Appropriate

MCDA is appropriate when: choosing between programme options with multiple relevant dimensions (outcomes, costs, feasibility, acceptability, equity, sustainability), decisions where stakeholders hold different values requiring explicit negotiation, planning decisions where trade-offs between objectives need transparency, resource allocation across programmes with different outcome types, policy decisions requiring audit trail showing how conclusions were reached, situations where monetising outcomes is inappropriate or contested, EU frameworks emphasising proportionality and multiple objectives (European Commission, 2021).

MCDA is less suitable when: simple decisions with single clear criterion (use simpler analysis), decisions already made where MCDA would be window dressing, contexts where stakeholders fundamentally disagree about what should be measured (resolve this first before MCDA), very small decisions where MCDA process costs exceed value of improved decision.

Key Principles

Comprehensive yet focused criteria: Criteria should capture all important dimensions, be distinct with minimal overlap, be measurable (quantitatively or qualitatively), be comprehensible to stakeholders, and number manageably (typically 5-12 criteria—more becomes unwieldy). Avoid criteria that are double-counting (correlates of same underlying dimension), unmeasurable even qualitatively, or universally satisfied by all options (no discriminating power).

Explicit transparent weighting: Weights reflect relative importance of criteria. The weighting process must be transparent and stakeholder-driven. Methods include: direct allocation (distribute 100 points across criteria), swing weighting (rank criteria by importance of moving from worst to best performance), pairwise comparison (compare criteria importance two at a time). Weights should reflect genuine stakeholder values, not analyst assumptions. Different stakeholder groups may weigh differently—this surfaces value conflicts requiring negotiation.

Evidence-based scoring: Score each alternative on each criterion using best available evidence: quantitative data where available, expert judgment where quantitative data lacking, stakeholder assessment where subjective criteria. Scoring should be systematic and documented. Avoid scoring based on preferences for alternatives—score based on evidence of actual performance.

Systematic sensitivity analysis: Test how conclusions change when varying weights, varying scores within plausible ranges, adding or removing criteria, using different stakeholder group weights. MCDA conclusions are robust when findings hold across sensitivity analyses. Conclusions changing dramatically with small weight changes indicate close decision requiring careful deliberation.

Facilitated stakeholder process: MCDA works best as a facilitation tool for deliberation, not mechanical calculation. Process forces explicit discussion of what matters (criteria identification), makes value conflicts visible (weighting), creates shared evidence base (scoring discussion), enables exploration of trade-offs (sensitivity analysis). Value lies in structured deliberation, not only in final scores.

Quality Standards

MCDA must meet minimum quality standards:

Comprehensive criteria: Criteria capture all important dimensions. Criteria are distinct (minimal overlap). Criteria are measurable. Number of criteria manageable (5-12 typical).

Stakeholder-driven weighting: Weighting process involves relevant stakeholders (decision-makers, programme staff, service users where appropriate). Weighting method transparent (direct allocation, swing weighting, pairwise comparison).

Evidence-based scoring: Scoring uses best available evidence. Data sources documented. Expert judgment process documented where used. Scoring systematic across all alternatives.

Documented process: Criteria development process documented. Weighting process documented. Scoring process documented. Stakeholder participants identified. Meeting notes maintained.

Sensitivity analysis: Key assumptions tested. Results reported showing robustness or sensitivity. Threshold analysis identifies tipping points.

Transparent reporting: MCDA process described clearly. Criteria and weights presented transparently. Scores and evidence presented. Final rankings reported with sensitivity analysis. Limitations acknowledged.

Common Errors

Error 1: Analyst-driven criteria and weights. MCDA value lies in surfacing stakeholder values. Analyst imposing criteria or weights defeats purpose. Solution: genuine stakeholder engagement in criteria identification and weighting, document divergent stakeholder perspectives, use MCDA as a facilitation tool not calculation exercise.

Error 2: Double-counting. Including highly correlated criteria inflates the importance of underlying dimension. Example: including "participant satisfaction," "participant engagement," and "staff-rated rapport" likely measures the same underlying relationship quality. Solution: check criteria correlations, combine or remove redundant criteria, ensure criteria capture distinct dimensions.

Error 3: Spurious precision. Reporting final scores to multiple decimal places (Alternative A: 7.2847) misleadingly precise given subjective weighting and scoring. Solution: present results as approximate (Alternative A: ~7.3), emphasise sensitivity analysis, acknowledge uncertainty.

Error 4: Ignoring process value. Treating MCDA as calculation producing answer misses main value—structured deliberation surfacing values and trade-offs. Solution: invest time in facilitated stakeholder process, use MCDA to structure discussion, value increased understanding of trade-offs even when stakeholders still disagree about the preferred option.

Implementation Requirements

Expertise needed: MCDA requires facilitation skills and understanding of multi-criteria analysis methods but not specialist technical expertise at basic level. The facilitator should have: group facilitation skills, MCDA method knowledge, ability to structure complex decisions, neutrality enabling honest deliberation. Small programmes can implement simple MCDA using trained internal staff. Medium-large programmes benefit from external facilitators avoiding power dynamics where managers facilitate staff or service user discussions.

Time investment: Simple MCDA: 3-5 days total including half-day stakeholder workshop for criteria identification and weighting, data collection and scoring, analysis and reporting. Comprehensive MCDA: 2-4 weeks including multiple stakeholder consultations, systematic data collection, formal scoring protocols, extensive sensitivity analysis.

Stakeholder engagement: Minimum: decision-makers who will use MCDA results, key staff implementing programmes, representation from affected stakeholders where appropriate. Ideal: diverse stakeholder perspectives including service users, front-line staff, managers, partners, funders enabling surfacing of value conflicts and negotiation of trade-offs.

Commissioning external specialists: When commissioning MCDA, specify: decision problem clearly, alternatives being evaluated, stakeholder groups to involve, timeline, deliverables. Provide facilitators with: programme descriptions for alternatives being compared, available data on alternative performance, stakeholder contact information, meeting logistics. Reference Annex B.5 technical specifications requiring compliance with MCDA protocols and quality standards.

Resources:

- **Annex B.5.1:** MCDA planning template
- **Annex B.5.2:** Criteria identification worksheet
- **Annex B.5.3:** Weighting protocols (direct allocation, swing weighting)
- **Annex B.5.4:** Performance scoring template
- **Annex B.5.5:** MCDA calculation spreadsheet
- **Annex B.5.6:** Results presentation templates
- **Annex B.5.7:** Quality assurance protocol

Integration with other methods: MCDA builds on Theory of Change identifying relevant criteria. MCDA incorporates Cost-Effectiveness Analysis results as one criterion (cost-effectiveness) among many. MCDA complements Cost-Benefit Analysis when a decision involves multiple objectives beyond economic efficiency. MCDA synthesises Stakeholder Feedback translating diverse perspectives into structured decision frameworks.

3.3.3 Social Return on Investment

Definition and Purpose

Social Return on Investment applies cost-effectiveness principles with stakeholder-driven outcome identification and valuation (Nicholls et al., 2012), creating accessible social value metrics for medium-scale programmes and social investment contexts. SROI provides stakeholder-driven social value measurement more accessible than comprehensive Cost-Benefit Analysis whilst more systematic than basic outcome monitoring.

SROI measures and accounts for social, environmental, and economic value creation, monetizing outcomes important to stakeholders, calculating SROI ratio expressing social value created per unit invested (e.g., €4.50 social value per €1 invested). SROI is particularly common in social investment contexts (impact bonds, social finance, venture philanthropy) and when funders explicitly request stakeholder-driven value assessment.

When SROI is Appropriate

SROI appropriate when: social investment context with investors requiring SROI evidence (social impact bonds, venture philanthropy, community interest companies), stakeholder engagement central to programme values (participatory programmes where beneficiary voice essential), multiple stakeholder groups experience outcomes requiring valuation (participants, families, communities, employers all affected), funder explicitly requests SROI (some foundations and impact investors specify SROI in grant requirements), outcomes resist conventional economic valuation but stakeholders can articulate value (dignity, voice, empowerment valued by stakeholders even when economists struggle with monetary equivalents).

SROI less appropriate when: stakeholders fundamentally object to monetisation exercise regardless of participatory process, no SROI requirement from funders (SROI more resource-intensive than basic outcome monitoring without proportionate analytical benefit unless specifically valued by funders), programme scale too small for SROI investment, outcomes require comprehensive economic valuation across multiple benefit streams justifying investment in Cost-Benefit Analysis (Section 3.4.1).

PRACTITIONER ALERT

Do Not Compare SROI Ratios Across Different Programmes

An SROI ratio of €5:€1 for Programme A does not mean it creates more social value than Programme B with an SROI ratio of €3:€1. SROI ratios depend heavily on which outcomes are included, which financial proxies are chosen, what impact adjustments are applied, and how stakeholders define "material" outcomes. Two competent SROI practitioners evaluating the same programme may produce different ratios based on legitimate methodological choices.

SROI is most useful for: understanding the types and scale of value a single programme creates, communicating social value to funders and investors in accessible terms, identifying which programme components generate most value, and tracking whether value creation improves over time within the same programme using consistent methods.

SROI should not be used for: league tables ranking programmes by ratio, competitive funding decisions comparing different programmes' SROI ratios, or performance benchmarking across organisations using different methodologies. If comparing programmes is the objective, Cost-Effectiveness Analysis (Section 3.3.1) with common outcome measurement is methodologically more appropriate.

Proportionality threshold: SROI appropriate for established programmes with annual budgets €500k+ where social investment contexts or funder requirements justify 1-2% of programme budget for evaluation. Programmes under €500k should focus on Cost-Effectiveness Analysis unless funders specifically require SROI.³

Key Principles

Stakeholder-driven throughout: Stakeholders identify material outcomes (not analyst-imposed), participate in financial proxy selection, validate impact adjustments, interpret findings. Stakeholder engagement is genuine, not tokenistic. Multiple stakeholder groups consulted including participants, families, staff, partners, community members. Materiality principle focuses on outcomes stakeholders consider significant.

³ The €500k threshold is indicative, derived from average SROI evaluation costs (€5,000–15,000) representing approximately 1–2% of programme budget. It is not a prescriptive cut-off; organisations should apply proportionality judgement relative to their own budget and funder requirements. Costs vary across EU member states and local quotes should inform any commissioning decision.

Theory of Change foundation: SROI requires a clear Theory of Change showing inputs, activities, outputs, outcomes, impacts for each stakeholder group. Theory of Change identifies who experiences what outcomes through which pathways. SROI monetises these outcomes using stakeholder-validated proxies.

Impact adjustments mandatory: Deadweight (what would have happened anyway without programme), attribution (contribution of other factors alongside programme), displacement (programme benefits some at expense of others), drop-off (outcomes decay over time). All four adjustments applied to all outcomes with evidence and rationale documented. SROI without rigorous impact adjustments severely overstates social value.

Quality Standards

SROI must meet minimum standards: Genuine stakeholder engagement documented (multiple groups participated, meaningful input not tokenistic consultation, stakeholder perspectives shaped outcome selection and valuation). Theory of Change explicitly links activities to outcomes for each stakeholder group. Material outcomes identified through stakeholder process (outcomes selected based on stakeholder-defined significance not analyst convenience). Financial proxies justified and stakeholder-validated (valuation approach documented, sources cited, stakeholders consulted on appropriateness). All four impact adjustments applied systematically to every outcome (deadweight, attribution, displacement, drop-off percentages specified with evidence/rationale). Sensitivity analysis conducted (key assumptions varied, SROI presented as range not point estimate). Transparent reporting (assumptions, limitations, uncertainties acknowledged, sufficient detail enabling verification). Independent assurance obtained for high-stakes SROI (particularly for social impact bonds, contested results, funder requirement).

Common Errors

Error 1: Tokenistic stakeholder engagement. Consulting stakeholders superficially whilst analysts make real decisions. Solution: genuine participatory process, stakeholder input demonstrably shapes outcome selection and valuation, document how stakeholder feedback influenced analysis.

Error 2: Omitting impact adjustments. Calculating SROI without deadweight/attribution/displacement/drop-off dramatically overstates value. Solution: apply all four adjustments systematically to every outcome, document evidence/rationale for adjustment percentages, conservative assumptions when evidence is weak.

Error 3: Cherry-picking outcomes. Including only positive outcomes whilst ignoring negative effects or unintended consequences. Solution: comprehensive outcome

identification through Theory of Change, explicit consideration of potential negative outcomes, transparent reporting of outcomes included/excluded with rationale.

Error 4: Spurious precision. Reporting SROI as exact figure (€4.37 per €1 invested) misleadingly precise given subjective assumptions. Solution: present SROI as range from sensitivity analysis, emphasize uncertainty, avoid single point estimates suggesting false precision.

PRACTITIONER ALERT

Two Common Pitfalls in SROI

Inflating your numbers:

SROI requires assigning monetary values to outcomes that don't have market prices. This is where things can go wrong. If your programme helps someone gain confidence, don't value that using the salary premium of a university degree — the proxy must match the actual change participants experienced, not an aspirational version of it. Similarly, don't count overlapping outcomes separately (e.g., valuing "better mental health," "reduced anxiety," and "improved emotional wellbeing" as three distinct benefits when they describe the same underlying change), and don't project benefits far into the future without accounting for the reality that effects fade over time.

The safeguard is straightforward: when in doubt, be conservative. Ask your stakeholders whether the proxy genuinely reflects their experience. Test what happens to your ratio when you use lower values. An SROI of €2.50:€1 that nobody can challenge is worth far more than €8:€1 that falls apart under scrutiny. Inflated SROI doesn't just produce a wrong number — it gives critics ammunition to dismiss social value measurement altogether.

Comparing ratios across different programmes:

An SROI of €5:€1 for Programme A does not mean it creates more social value than Programme B at €3:€1. The ratios depend on which outcomes were included, which monetary proxies were chosen, what adjustments were applied, and how broadly "value" was defined. Two competent practitioners evaluating the same programme may legitimately produce different ratios.

SROI is useful for understanding the value your programme creates, communicating that value to funders, and tracking whether it improves over time using consistent methods. It should not be used for league tables, competitive funding comparisons, or cross-organisation benchmarking. If comparing programmes is the goal, Cost-Effectiveness Analysis (Section 3.3.1) with a common outcome measure is the appropriate method.



Implementation Requirements

Expertise needed: SROI-accredited practitioners (Social Value International accreditation or equivalent), stakeholder facilitation expertise, experience with participatory valuation methods, understanding of impact adjustments and sensitivity analysis. Check credentials carefully as SROI quality varies substantially across practitioners.

Typical investment: 1-3 months for established programmes. Budget requirements vary substantially across EU member states. Commissioners should obtain local quotes; SROI is typically more resource-intensive than Cost-Effectiveness Analysis but less so than full Cost-Benefit Analysis.

Data requirements: Outcome data for all material stakeholder groups. Financial proxy databases (Global Value Exchange primary resource for EU contexts). Stakeholder access for consultation. See Chapter 2 for data governance.

Commissioning SROI: Require SROI accreditation or equivalent credentials. Specify genuine stakeholder engagement requirements (not analyst-driven). Require comprehensive impact adjustments. Require sensitivity analysis presenting results as range. Reference Annex B.6.2 requires compliance with SROI technical specifications and Social Value International principles.

Resources: Annex B.6.2 SROI technical specifications and quality standards. Case Study 4 Octavia Foundation Employment Programme SROI demonstration. Social Value International Principles (global standards). Global Value Exchange (financial proxy database).

Integration with other methods: SROI builds on Theory of Change (Section 3.2.1) identifying outcomes for stakeholder groups. SROI incorporates Stakeholder Feedback (Section 3.2.3) throughout the process. SROI complements CBA (Section 3.4.1) when stakeholder-driven valuation is preferred over analyst-driven economic valuation.

Resources:

- Annex B.6.1: SROI planning template and materiality assessment
- Annex B.6.2: Stakeholder engagement protocols for SROI
- Annex B.6.3: Financial proxy selection and validation framework
- Annex B.6.4: Impact adjustment calculation templates (deadweight, attribution, displacement, drop-off)
- Annex B.6.5: SROI calculation spreadsheet with sensitivity analysis
- Annex B.6.6: SROI reporting templates and assurance protocols
- Case Study 4: Octavia Foundation Employment Programme SROI demonstration

Integration with other methods:

SROI builds on Theory of Change (Section 3.2.1) identifying outcomes for each stakeholder group and establishing causal pathways requiring monetization. SROI uses Outcome Monitoring (Section 3.2.2) data providing outcome quantities for valuation. SROI incorporates Stakeholder Feedback (Section 3.2.3) throughout the participatory process of outcome identification, proxy validation, and results interpretation. SROI sits between Cost-Effectiveness Analysis (Section 3.3.1) and Cost-Benefit Analysis (Section 3.4.1) in sophistication—more comprehensive than CEA in capturing multiple stakeholder values whilst more accessible than CBA in resource requirements and stakeholder participation. SROI complements Multi-Criteria Decision Analysis (Section 3.3.2) when social value quantification supplements structured multi-criteria frameworks.

3.4 ADVANCED METHODS (Comprehensive Valuation and Causality)

Advanced Methods provide definitive proof of causal impact or comprehensive economic valuation for large-scale programmes or high-stakes policy decisions. These methods require specialist expertise—health economists, trial methodologists, causal inference specialists, skilled qualitative researchers—and substantial investment varying by method and context. Programmes should commission external specialists rather than attempt DIY implementation of methods requiring technical skills beyond typical programme staff competence.

Commissioning External Specialists for Advanced Methods

Advanced Methods require specialist expertise and cannot be implemented reliably using internal programme capacity alone. Programmes should commission external specialists—academic research teams, government analytical services, specialist evaluation consultancies—rather than attempt implementation requiring advanced technical skills.

Minimum specialist qualifications vary by method: CBA requires economists or public policy analysts with postgraduate qualifications and demonstrated CBA experience, published evaluation work including peer-reviewed journal articles, experience evaluating social services. Quasi-Experimental Design requires causal inference specialists with demonstrated expertise in statistical matching methods. Randomised Controlled Trials require trial methodologists, statisticians, and research infrastructure. Realist Evaluation requires skilled qualitative researchers with complexity methods expertise.

Commissioning requirements: Define evaluation objectives clearly (what decisions will evaluation inform, what questions must be answered, what certainty level required, what

timeline constraints exist). Specify programme context comprehensively (programme description, population served (target group), implementation model, existing data availability, stakeholder access, policy context). Request credentials and work samples (CVs demonstrating relevant expertise, prior evaluation examples using proposed methods, references, publications where applicable). Reference technical standards (evaluations must meet quality standards in Annexes B.7-B.8, commissioners provide these specifications to evaluators requiring compliance with analytical protocols and reporting requirements). Budget realistically (under-resourced evaluations produce unreliable results, obtain cost estimates from multiple qualified specialists).

Quality assurance: Independent review for high-stakes evaluations (commission independent quality review by second specialist not involved in primary evaluation—particularly important for CBA where second health economist reviews valuation choices, QED/RCT where second statistician reviews causal identification strategy). Transparent documentation (evaluators must document all assumptions, data sources, analytical choices, limitations enabling replication by independent analysts). Stakeholder engagement throughout (even when commissioning technical specialists, programme staff and stakeholders remain engaged providing programme knowledge, facilitating data access, interpreting findings for practical relevance). Ethical governance (research ethics review required for RCTs, typically required for QED involving vulnerable populations, recommended for all Advanced Methods involving primary data collection, GDPR compliance for all data handling).

Resources: Annex B.7 (CBA technical specifications and quality standards), Annex B.8 (QED, RCT, Realist Evaluation technical specifications and quality standards). These Annexes provide analytical protocols, quality standards, and reporting requirements. Commissioners reference these specifications when engaging evaluators, requiring compliance with documented standards.

Reality check: Most small-medium organisations cannot afford Advanced Methods unless part of multi-site evaluation sharing costs, funder provides dedicated evaluation grant, or pro-bono academic partnership secured. Organisations lacking these conditions should focus on excellent implementation of Foundation and Intermediate Methods rather than under-resourced Advanced Methods producing unreliable results.

3.4.1 Cost-Benefit Analysis

Definition and Purpose

Cost-Benefit Analysis monetises all costs and benefits of interventions determining whether total benefits exceed total costs (Boardman et al., 2018). CBA converts all effects into monetary terms enabling comparison across programmes with entirely different outcome types. CBA is the gold standard for major policy decisions forming

core of UK Treasury Green Book (HM Treasury, 2022) and EU Better Regulation Guidelines (European Commission, 2021).

CBA systematically identifies, quantifies, and monetises all costs and benefits from societal perspective, discounts future values to present terms, and calculates whether net benefits are positive. Results expressed as Benefit-Cost Ratio (BCR = total benefits / total costs, where BCR >1.0 indicates benefits exceed costs) or Net Present Value (NPV = total benefits - total costs, where positive NPV indicates net social value creation).

When Cost-Benefit Analysis is Appropriate

CBA appropriate when: major policy decisions requiring comprehensive economic assessment (programmes affecting large populations, substantial public investment decisions, policy mandates affecting multiple sectors), scaling decisions requiring evidence economic benefits justify expansion costs, cross-sector comparison needed (comparing housing programme to employment programme to health intervention), funder requirements explicitly request CBA (EU Better Regulation for major policies, What Works Centres for evidence synthesis, major foundations for impact bonds), monetisation feasible and acceptable for primary outcomes (employment, health service use, criminal justice outcomes have established monetary values—dignity, justice, community cohesion resist monetisation).

CBA inappropriate or infeasible when: programme scale small relative to CBA investment (CBA typically requires substantial specialist investment disproportionate to small programme decisions), primary outcomes resist monetary valuation and stakeholders object to monetisation attempts (forcing monetary values on human dignity, justice, rights undermines credibility), causal attribution unclear (CBA shows association between programme and monetised outcomes but cannot prove causation without comparison group designs—Sections 3.4.2, 3.4.3), data insufficient for reliable benefit quantification (incomplete outcome data, inability to track long-term effects, missing cost data).

Key Principles

Societal perspective: CBA measures value to society not organisations. Include all costs regardless who pays (programme costs, participant costs, public sector costs across departments, societal costs). Include all benefits regardless who receives (earnings gains, health improvements, criminal justice savings, reduced service use, wellbeing improvements, productivity gains). Narrower perspectives (programme, funder, public sector) may be reported alongside societal perspective but societal perspective provides comprehensive economic assessment.

Comprehensive benefit identification: Systematic identification of all significant benefits using Theory of Change as framework (direct participant benefits, spillover



effects to families/communities, reduced costs to public services, productivity gains to employers, societal benefits from reduced crime/improved health/stronger communities). Benefits commonly overlooked include: reduced informal care burden on families, avoided future costs prevented by early intervention, spillover benefits to non-participants, option value of maintaining service availability, existence value society places on helping vulnerable populations.

Monetary valuation methods: Benefits monetised using: market prices for goods/services with established markets, revealed preference where behaviour reveals value (will-being to pay for gym membership reveals value of fitness), stated preference where surveys ask hypothetical willingness to pay (use cautiously as stated often exceeds actual willingness), cost savings for reduced service use (health service unit costs, criminal justice costs, education costs), human capital approach valuing productivity gains through earnings, wellbeing valuation assigning monetary values to life satisfaction changes (based on studies estimating income equivalent of wellbeing changes), transfer values from prior CBAs in similar contexts (most pragmatic approach when direct valuation infeasible). Valuation method choice requires justification and sensitivity analysis.

Discounting future values: Future costs and benefits discounted to present value reflecting time preference and opportunity cost. EU standard: 3% social discount rate per EU Better Regulation Vademecum (European Commission, 2021). UK standard: 3.5% discount rate (HM Treasury, 2022). Apply consistently to costs and benefits. Discount rate choice significantly affects results when benefits occur long after costs—sensitivity analysis testing alternative discount rates (1.5%, 3%, 5%) shows robustness.

Distributional analysis: Beyond aggregate BCR/NPV, assess who gains and who loses. Distributional analysis essential for equity concerns: do benefits accrue primarily to initially disadvantaged (progressive) or advantaged (regressive) groups, do certain groups bear costs whilst others receive benefits, are trade-offs between aggregate efficiency and distributional equity. EU Better Regulation Guidelines (European Commission, 2021) require distributional analysis alongside aggregate economic assessment.

Uncertainty management: CBA involves substantial uncertainty in benefit quantification, monetary valuation, long-term projections, discount rate choice. Address through: sensitivity analysis varying key parameters, scenario analysis exploring optimistic/pessimistic assumptions, probabilistic analysis where data support Monte Carlo simulation, presenting results as ranges not point estimates, threshold analysis identifying parameter values at which BCR crosses 1.0.

Quality Standards

CBA must meet minimum quality standards: Societal perspective applied consistently (all costs and benefits counted regardless who pays/receives). Comprehensive benefit identification using Theory of Change as framework (benefits systematically identified not cherry-picked). Justified monetary valuation methods (valuation approach specified for each benefit category with sources cited, alternative valuations tested in sensitivity analysis). Discounting applied correctly (3% EU or 3.5% UK standard, applied consistently, sensitivity tested). Distributional analysis conducted (who gains/loses, equity implications assessed). Sensitivity analysis comprehensive (key uncertainties tested systematically, results presented as ranges, thresholds identified). Non-monetised outcomes reported explicitly (benefits resisting monetisation presented alongside BCR/NPV not hidden, decision-makers see full value created). Transparent documentation (methods enabling replication, assumptions stated explicitly, data sources specified, limitations acknowledged). Independent quality assurance for high-stakes CBAs (second health economist reviews valuation choices and analytical approach).

Common Errors

Error 1: Forcing monetisation of contested values. Attempting to monetise human dignity, justice, rights, community cohesion when stakeholders fundamentally object undermines credibility. Solution: present non-monetised benefits explicitly alongside BCR/NPV, acknowledge some values that resist monetisation, provide qualitative description of non-monetised impacts, ensure decision-makers see full value including what cannot be monetised.

Error 2: Incomplete benefit identification. Cherry-picking easily monetisable benefits whilst omitting harder-to-value impacts biases results downward. Solution: use Theory of Change systematically identifying all significant benefits, explicitly address each outcome pathway, document benefits excluded with justification, stakeholder consultation ensures no major benefits overlooked.

Error 3: Short time horizons missing long-term benefits. Many social service benefits (early intervention preventing later problems, education improving life course, health promotion avoiding chronic disease) emerge over decades. Too-short time horizons understate benefits. Solution: justify time horizon based on when benefits materialise, test longer time horizons in sensitivity analysis, acknowledge if evaluation occurs before full benefits emerged, reference long-term evidence from other studies supporting benefit persistence.

Error 4: Treating CBA as causal evaluation. CBA shows association between programme and monetized outcomes but cannot prove causation without comparison group controlling for selection. Solution: be transparent that CBA assumes outcomes



attributable to programme (assumption requiring testing through QED/RCT if causal attribution essential for decision), combine CBA with causal evaluation methods where stakes justify investment in both.

Implementation Requirements

Expertise needed: CBA requires economists or public policy analysts with postgraduate qualifications and demonstrated CBA experience, published evaluation work including peer-reviewed journal articles, and/or experience evaluating social services.

Typical investment: Time: 2-4 months for evaluation (data collection 2-4 weeks, analysis 4-8 weeks, reporting 2-4 weeks). Longer for complex programmes with multiple outcome pathways or long-term benefit streams requiring projections. Costs vary substantially by context, available data, and analysis complexity.

Data requirements: Cost data (comprehensive cost accounting across all categories, participant numbers, service intensity). Outcome data (measured outcomes for all significant benefit categories identified in Theory of Change, adequate sample sizes, follow-up periods capturing benefit emergence). Comparator data where possible (comparison group strengthens causal attribution, administrative data on service use by non-participants provides context). See Chapter 2 for data governance standards.

Commissioning CBA: Specify comprehensive benefit identification required (provide Theory of Change, require systematic review of all outcome pathways). Require societal perspective and distributional analysis. Require comprehensive sensitivity analysis. Provide available data (cost records, outcome data, administrative data access where available). Establish a realistic timeline for quality work. Reference Annex B.6 requires compliance with CBA technical specifications and quality standards.

Resources:

- Annex B.7: CBA technical specifications and quality standards
- Case Study 3: Housing First Manchester CBA demonstration

Technical standards references:

- EU Better Regulation standards (European Commission, 2021)
- UK Treasury Green Book CBA guidance (HM Treasury, 2022)

Integration with other methods: CBA builds on Theory of Change (Section 3.2.1) identifying benefit categories. CBA complements CEA (Section 3.3.1) when comprehensive economic assessment is needed beyond efficiency comparison. CBA combines with QED/RCT (Sections 3.4.2, 3.4.3) when causal attribution is essential alongside economic valuation.

3.4.2 Quasi-Experimental Design

Definition and Purpose

Quasi-Experimental Design uses comparison groups to estimate counterfactual outcomes (what would have happened without programme) whilst lacking random assignment (Shadish et al., 2002). QED employs statistical methods controlling for pre-existing differences between participants and comparisons, establishing whether programmes cause observed outcomes.

QED answers: "Did this programme cause these outcomes or would they have occurred anyway?" QED provides stronger causal evidence than outcome monitoring alone but carries residual uncertainty about unmeasured confounding compared to randomised controlled trials.

When QED is Appropriate

QED appropriate when: causal attribution essential for scaling decisions or policy mandates, random assignment infeasible or unethical (universal entitlement, urgent need, ethical objections to withholding services), comparison group identifiable (waiting list, adjacent geographic area, similar programme in different location, administrative data on non-participants), adequate sample sizes available (typically 200+ participants and 200+ comparisons minimum for reliable matching), baseline data available enabling statistical matching or difference-in-differences analysis, resources available for specialist evaluation (QED requires causal inference expertise and substantial investment).

QED inappropriate when: no plausible comparison group identifiable, sample sizes too small for adequate statistical power, baseline data insufficient for matching or controlling confounds, random assignment feasible making RCT (Section 3.4.3) more rigorous option, resources insufficient for quality implementation (poor-quality QED worse than transparent outcome monitoring—under-resourced QED produces misleading causal claims).

Key Principles

Comparison group equivalence: QED validity depends on comparisons being similar to participants on all factors affecting outcomes. Statistical methods address observable differences (propensity score matching, regression adjustment, difference-in-differences). Residual selection bias is possible if groups differ on unmeasured factors. Equivalence assessment critical: compare groups on baseline characteristics, assess overlap in propensity score distributions, test sensitivity to unobserved confounding, acknowledge limitations honestly.

Adequate statistical power: QED requires sufficient sample sizes detecting meaningful effect sizes. Underpowered studies produce unreliable results. Minimum 200 participants and 200 comparisons typical for adequately powered QED. Larger samples needed if effect sizes are small or outcome variance high. Power calculations should inform sample size planning.

Quality Standards

QED must meet minimum standards: Comparison group justified and well-matched (comparison source specified with rationale, equivalence demonstrated on baseline characteristics, overlap in distributions adequate for matching). Statistical method appropriate (propensity score matching, difference-in-differences, regression discontinuity, or instrumental variables implemented correctly with diagnostics reported). Adequate statistical power (sample sizes sufficient for intended inferences, power calculations documented). Sensitivity analysis conducted (robustness tested to matching approach, covariate specification, unobserved confounding assumptions). Transparent reporting (sufficient detail enabling replication, comparison group source clear, matching approach documented, limitations acknowledged). Independent quality assurance (second statistician reviews causal identification strategy and analytical approach).

Common Errors

Error 1: Poor comparison group selection. Comparisons fundamentally different from participants producing biased results. Solution: careful comparison group selection, rigorous equivalence assessment, transparent reporting of differences, sensitivity analysis, honest acknowledgment of limitations.

Error 2: Inadequate statistical power. Too-small samples unable to detect plausible effect sizes. Solution: power calculations informing sample size, realistic effect size expectations, acknowledgment when power insufficient, avoid causal claims from underpowered studies.

Error 3: Treating QED as definitive proof. QED provides strong evidence but cannot eliminate all threats to causal inference. Solution: transparent limitations discussion, sensitivity analysis testing unobserved confounding assumptions, triangulation with other evidence where possible.

Implementation Requirements

Expertise needed: Causal inference specialists—econometricians, quantitative social scientists, evaluation researchers with demonstrated expertise in propensity score matching, difference-in-differences, regression discontinuity, or instrumental variables.

Check credentials: peer-reviewed publications using these methods, prior evaluations demonstrating causal inference competence, recommendations from credible sources.

Typical investment: Time: 6-12 months (comparison group identification and recruitment 2-3 months, data collection 3-6 months, analysis 2-3 months, reporting 1 month). Costs vary substantially by comparison group accessibility, data availability, sample size requirements.

Data requirements: Outcome data for participants and comparisons measured identically. Baseline data enabling matching (demographics, baseline outcomes, selection factors). Adequate sample sizes (200+ each group minimum). See Chapter 2 for data governance.

Commissioning QED: Specify causal questions clearly. Describe available comparison group sources. Provide available data (participant data, potential comparison data or access). Require power calculations. Reference Annex B.8.1 requiring compliance with QED technical specifications and quality standards.

Resources:

- Annex B.8.1: QED technical specifications and quality standards
- EU Better Regulation impact evaluation guidance (European Commission, 2021)

Integration with other methods: QED builds on Theory of Change (Section 3.2.1) specifying causal pathways to test. QED uses Outcome Monitoring (Section 3.2.2) data. QED combines with CBA (Section 3.4.1) when economic valuation is needed alongside causal evidence.

3.4.3 Randomised Controlled Trials

Definition and Purpose

Randomised Controlled Trials represent the gold standard for causal inference (Shadish et al., 2002), using random assignment creating equivalent treatment and control groups thereby eliminating selection bias. RCTs answer definitively: "Does this programme cause observed outcomes?"

RCTs randomly assign eligible individuals or clusters to intervention or control, then compare outcomes between groups. Randomisation ensures groups equivalent in expectation on all characteristics—measured, unmeasured, unknown—isolating programme effects from confounding factors.

When RCT is Appropriate

RCT appropriate when: definitive causal evidence required for scaling decisions, policy mandates, major investment, ethical equipoise exists (genuine uncertainty whether programme helps making randomisation ethically acceptable), programme mature and standardised (stable implementation not still being refined), sufficient sample size available (typically 300+ individuals minimum for adequate power), resources available (RCTs most expensive evaluation method), timeframe adequate (18-36 months from design to results often 24-48 months for complex trials), funders and policymakers value experimental evidence (What Works Centres, major foundations, government innovation funds, EU impact evaluation requirements for HORIZON innovation projects).

RCT inappropriate when: unethical to withhold intervention (life-saving, urgent need, legally mandated services), no equipoise (clear evidence intervention works making randomisation unjustifiable), universal entitlement (everyone eligible must be served immediately), pilot or innovation stage (programme still being refined premature for definitive evaluation), small scale (insufficient sample size for adequate power), implementation variation too great (no standardised intervention to test), budget insufficient (poor-quality RCT worse than well-conducted QED).

Key Principles

Methodological rigour: Random assignment properly implemented (concealed allocation, prevention of selection bias, documented randomisation procedures). Adequate sample size (power calculations ensuring sufficient statistical power). Outcome measurement blinded where possible (assessors unaware of treatment assignment). Intention-to-treat analysis (participants analysed as randomised regardless of actual participation). Complete reporting following CONSORT guidelines.

Ethical requirements: Research ethics approval mandatory. Informed consent procedures protecting vulnerable populations. Equipoise established (genuine uncertainty whether intervention helps). Control conditions ethically acceptable (may receive standard care, delayed intervention, or nothing depending on context). Data monitoring ensures participant safety. Trial registration in the public registry.

Quality Standards

RCT must meet minimum standards: Proper randomisation (concealed allocation preventing selection bias, randomisation sequence adequately generated, documentation enabling verification). Adequate statistical power (sample size sufficient detecting meaningful effects, power calculations documented). Appropriate outcome measurement (validated instruments where available, assessment blinded to assignment where feasible, follow-up adequate for outcome emergence). Intention-to-treat analysis conducted (participants analysed as randomised). Complete reporting per

CONSORT guidelines (flow diagram showing participant progress, attrition documented, baseline equivalence demonstrated, effect estimates with confidence intervals) (Schulz et al., 2010). Research ethics approval obtained. Trial registered prospectively. Independent data monitoring for complex trials.

Common Errors

Error 1: Inadequate concealment. Selection bias when allocation is predictable. Solution: proper concealed allocation procedures, independent randomisation service, documentation enabling verification.

Error 2: Underpowered trials. Too-small samples unable to detect plausible effects. Solution: power calculations informing sample size, realistic effect size expectations, adequate recruitment.

Error 3: High attrition undermining randomisation. Differential loss to follow-up introduces bias. Solution: strategies maximising retention, attrition analysis, sensitivity analysis, honest acknowledgment when attrition compromises validity.

Implementation Requirements

Expertise needed: Trial methodologists with RCT design experience, statisticians with clinical trials expertise, research coordinators, data managers. Academic research teams typically require given infrastructure needs (ethics approval, trial registration, data monitoring).

Typical investment: Time: 18-36 months (design and ethics 3-6 months, recruitment 6-12 months, intervention and follow-up 6-18 months, analysis and reporting 3-6 months). Costs vary substantially by scale and complexity. Among the most expensive evaluation methods.

Data requirements: Outcome measures for all participants. Baseline data. Process data documenting intervention delivery and fidelity. See Chapter 2 for data governance. Ethics approval mandatory.

Commissioning RCT: Specify causal question. Describe intervention and control conditions. Provide sample size information. Establish a realistic timeline. Budget appropriately. Reference Annex B.8.2 requiring compliance with RCT technical specifications and CONSORT reporting standards (Schulz et al., 2010) - RCT reporting standards.

Resources:

- Annex B.8: RCT technical specifications and quality standards
- CONSORT Statement (Schulz et al., 2010) - RCT reporting standards

- Trial registries: ISRCTN (www.isrctn.com), ClinicalTrials.gov (www.clinicaltrials.gov)
- MRC Process Evaluation guidance (Moore et al., 2015)

Integration with other methods: RCT builds on Theory of Change (Section 3.2.1) specifying mechanisms to test. RCT combines with Realist Evaluation (Section 3.4.4) understanding how context shapes outcomes. RCT combines with CBA (Section 3.4.1) when economic valuation is needed alongside causal evidence.

3.4.4 Realist Evaluation

Definition and Purpose

Realist Evaluation addresses: "What works, for whom, in what circumstances, and how?" (Pawson & Tilley, 1997). Rather than testing whether programmes work, realist evaluation unpacks causal mechanisms and contextual conditions producing outcomes. Particularly powerful for understanding why programmes succeed in some contexts but fail in others.

Realist evaluation is a theory-driven approach examining how programme mechanisms interact with contexts generating outcomes, typically expressed as Context-Mechanism-Outcome configurations. Incorporates Contribution Analysis (Mayne, 2012) for assessing causation in complex multi-actor environments where attribution to single programmes is impossible.

When Realist Evaluation is Appropriate

Realist evaluation appropriate when: complex interventions with multiple components and pathways, outcomes depend substantially on context (implementation setting, participant characteristics, local resources, policy environment), understanding HOW and WHY programmes work as important as WHETHER they work, programme operates in complex multi-actor environment making simple attribution impossible, scaling or adaptation requires understanding which mechanisms essential and which contexts enabling, qualitative depth and contextual understanding valued alongside quantitative patterns.

Realist evaluation less suitable when: simple interventions with clear mechanisms in standard contexts (outcome monitoring sufficient), definitive yes/no causal conclusion required (use RCT/QED), programme operates independently in controlled setting (realist complexity methods unnecessary), resources insufficient for quality qualitative research (realist evaluation requires skilled qualitative researchers and substantial time investment).

Key Principles

Theory-driven: Realist evaluation develops and tests theories about Context-Mechanism-Outcome configurations. Theory specifies: in what contexts (C), which mechanisms fire (M), producing which outcomes (O) for which participants. Multiple CMO configurations typically needed explaining variation in outcomes. Theory refined iteratively through data collection and analysis.

Mechanism focus: Mechanisms are underlying processes by which programme activities produce outcomes—not activities themselves but how activities trigger change (reasoning, resources enabling or constraining behaviour, relationships, information, capability building, motivation). Understanding mechanisms enables identifying which programme components are essential versus peripheral.

Context sensitivity: Outcomes depend on context—implementation quality, participant readiness, local resources, policy environment, cultural factors. Realist evaluation systematically examines how context shapes which mechanisms fire for which participants. Enables understanding transferability: which contexts enable programme success, which require adaptation, which make programmes ineffective.

Contribution not attribution: Contribution Analysis integrated within realist approach acknowledges programmes contribute alongside other factors in complex multi-actor environments. Contribution Analysis builds evidence-based cases for programme contribution without claiming exclusive attribution. Appropriate for policy evaluation, systems-level interventions, multi-partner programmes where multiple factors produce outcomes simultaneously.

Quality Standards

Realist evaluation must meet minimum standards: Explicit theory development (CMO configurations specified clearly, theory grounded in evidence and stakeholder knowledge, theory testable through data collection). Appropriate data collection (data sources adequate testing theory, purposive sampling captures contextual variation, qualitative methods rigorous). Systematic analysis (CMO configurations tested against data, rival explanations considered, theory refined based on evidence). Context-mechanism interaction demonstrated (shows HOW mechanisms work differently in different contexts, explains outcome variation, enables transferability assessment). Contribution claim supported (evidence programme contributed to outcomes, rival explanations addressed, claims proportionate to evidence strength). Transparent reporting (sufficient detail enabling assessment, limitations acknowledged, theory evolution documented).

Common Errors

Error 1: Describing not explaining. Reporting what happened without unpacking mechanisms and contexts. Solution: explicit mechanism identification, theory testing against data, explanation of how and why outcomes are produced.

Error 2: Ignoring context variation. Treating context as background not active ingredient. Solution: systematic context analysis, demonstration of context-mechanism interaction, explanation of outcome variation by context.

Error 3: Over-claiming causation. Making attribution claims inappropriately given complexity. Solution: contribution framing, rival explanations considered, proportionate claims, transparency about what can and cannot be concluded.

Implementation Requirements

Expertise needed: Skilled qualitative researchers with demonstrated expertise in realist methods, CMO configuration development, theory-driven evaluation, complexity-informed approaches. Check credentials: prior realist evaluations, publications demonstrating realist thinking, recommendations from credible sources.

Typical investment: Time: 12-24 months (theory development 2-3 months, data collection 6-12 months, analysis 3-6 months, reporting 1-3 months). Costs vary by scale and data collection intensity.

Data requirements: Qualitative data (interviews, focus groups, observations, documents) enabling mechanism and context exploration. Quantitative outcome patterns where available providing context. Theory of Change as a starting point. See Chapter 2 for data governance.

Commissioning Realist Evaluation: Specify interest in understanding how and why programmes work not only whether they work. Provide Theory of Change or programme theory. Describe contextual variation of interest. Reference Annex B.8.3 requiring compliance with realist evaluation technical specifications and quality standards.

Resources:

- Annex B.8.3: Realist evaluation and contribution analysis technical specifications and quality standards
- Pawson & Tilley (1997) - foundational realist evaluation methodology
- RAMESES reporting standards (Wong et al., 2016) - quality standards for realist evaluations
- Mayne (2012) - Contribution Analysis framework

Integration with Other Methods: Realist evaluation builds on Theory of Change (Section 3.2.1) as initial programme theory specifying causal mechanisms and contextual assumptions requiring testing. Realist evaluation synthesises Outcome Monitoring

(Section 3.2.2) data identifying outcome patterns across contexts and Stakeholder Feedback (Section 3.2.3) revealing participant experiences of mechanisms and contextual factors. Realist evaluation complements RCT (Section 3.4.3) by explaining why trials produce observed results, which mechanisms are fired in which contexts, and how findings transfer to different settings. Realist evaluation complements QED (Section 3.4.2) when comparison group designs show whether programmes work whilst realist methods explain how and why they work. Realist evaluation informs CBA (Section 3.4.1) by identifying which benefit pathways operate in which contexts, enabling more accurate benefit quantification and projection.

3.5 Commissioning External Specialists for Intermediate-Advanced Methods

Intermediate and Advanced Methods typically require specialist expertise beyond internal programme capacity. Programmes should commission external specialists—academic research teams, government analytical services, or specialist evaluation consultancies—rather than attempt DIY implementation of methods requiring technical skills beyond typical programme staff competence.

Minimum specialist qualifications vary by method:

Social Return on Investment: SROI-accredited practitioners (Social Value International accreditation or equivalent), stakeholder facilitation expertise, experience with participatory valuation methods.

Cost-Benefit Analysis: Economists or public policy analysts with sufficient qualification, experience evaluating social services, familiarity with EU Better Regulation Guidelines and UK Treasury Green Book standards.

Quasi-Experimental Design: Causal inference specialists—econometricians, quantitative social scientists, evaluation researchers with demonstrated expertise in propensity score matching, difference-in-differences, regression discontinuity, or instrumental variables.

Randomised Controlled Trials: Trial methodologists with RCT design experience, statisticians with clinical trials expertise, research coordinators, data managers. Academic research teams typically require given infrastructure needs (ethics approval, trial registration, data monitoring).

Realist Evaluation: Skilled qualitative researchers with demonstrated expertise in realist methods, context-mechanism-outcome configuration development, theory-driven evaluation, complexity-informed approaches.

Commissioning requirements:

Define evaluation objectives clearly: What decisions will evaluation inform? What questions must be answered? What level of certainty is required? What timeline constraints exist?

Specify programme context comprehensively: Programme description, population served (target group), implementation model, existing data availability, stakeholder access arrangements, relevant policy context.

Request credentials and work samples: CVs demonstrating relevant expertise, examples of prior evaluations using proposed methods, references from previous clients, publications where applicable.

Reference technical standards: Evaluations must meet quality standards specified in Annexes B.6-B.8. Commissioners should provide these technical specifications to prospective evaluators, requiring compliance with analytical protocols, reporting requirements, and quality standards detailed therein.

Budget realistically: Advanced Methods require substantial investment. Under-resourced evaluations produce unreliable results. Obtain cost estimates from multiple qualified specialists. The budget should cover specialist time, data collection costs, participant incentives where applicable, research ethics review if required, quality assurance processes.

Quality assurance requirements:

Independent review: For high-stakes evaluations, commission independent quality review by a second specialist not involved in primary evaluation. Particularly important for CBA (second health economist reviews valuation choices), QED/RCT (second statistician reviews causal identification strategy).

Transparent documentation: Evaluators must document all assumptions, data sources, analytical choices, limitations. Documentation should enable replication by independent analysts given access to the same data.

Stakeholder engagement: Even when commissioning technical specialists, programme staff and stakeholders must remain engaged throughout evaluation—providing programme knowledge, facilitating data access, interpreting findings for practical relevance.

Ethical governance: Research ethics review required for RCTs, typically required for QED involving vulnerable populations, recommended for all Advanced Methods involving primary data collection. Ensure GDPR compliance for all data handling.

Technical specifications appear in Annexes:

- **Annex B.6:** SROI technical specifications and quality standards
- **Annex B.7:** CBA technical specifications and quality standards
- **Annex B.8:** QED, RCT, and Realist Evaluation technical specifications and quality standards

These Annexes provide analytical protocols, quality standards, and reporting requirements. Commissioners should reference these specifications when engaging evaluators, requiring compliance with documented standards.

Reality check: Most small-medium organisations cannot afford these methods unless part of multi-site evaluation sharing costs, funder provides dedicated evaluation grant, or pro-bono academic partnership secured. Organisations lacking these conditions should focus on excellent implementation of Foundation and Intermediate Methods rather than under-resourced Advanced Methods producing unreliable results.

Chapter 4:

EU Integration and regulatory compliance



CHAPTER 4: EU INTEGRATION AND REGULATORY COMPLIANCE

4.1 EU Evaluation Mandates and Method Alignment

European programmes operate under specific evaluation requirements established by funding frameworks and regulatory guidelines. This chapter maps evaluation methods from this Green Book to EU mandates, enabling compliance whilst maintaining methodological rigour.

HORIZON Europe Requirements

Mandatory for all HORIZON projects: Theory of Change (Section 3.2.1) required in proposal stage demonstrating causal logic. Outcome Monitoring (Section 3.2.2) continues throughout project tracking indicators specified in Grant Agreement. Stakeholder Feedback (Section 3.2.3) regular consultation documenting engagement with relevant stakeholders. Data Management Plan (Chapter 2, Annex A.1) meeting HORIZON requirements including FAIR principles (Findability, Accessibility, Interoperability, Reusability).

For impact assessment: Quasi-Experimental Design (Section 3.4.2) or Randomised Controlled Trials (Section 3.4.3) establishing causal attribution where feasible and proportionate. Realist Evaluation (Section 3.4.4) understanding how programmes work in complex multi-partner contexts where attribution is contested. Multi-method approaches combining quantitative indicators with qualitative case studies.

For scaling and sustainability: Cost-Effectiveness Analysis (Section 3.3.1) or Cost-Benefit Analysis (Section 3.4.1) demonstrating value for money. Realist Evaluation (Section 3.4.4) understanding transferability conditions across contexts.

EU Better Regulation Requirements

Impact Assessment: Cost-Benefit Analysis (Section 3.4.1) or Multi-Criteria Decision Analysis (Section 3.3.2) for major policy options. CBA preferred for policies with substantial budgets or affecting large populations. MCDA appropriate when monetization inappropriate or multiple competing objectives require explicit trade-off analysis.

Evaluation: Quasi-Experimental Design (Section 3.4.2) or Randomised Controlled Trials (Section 3.4.3) for causal claims about programme effectiveness. Outcome Monitoring (Section 3.2.2) minimum for ongoing performance tracking.



Stakeholder Consultation : Stakeholder Feedback (Section 3.2.3) documenting genuine engagement. Theory of Change (Section 3.2.1) developed through participatory process.

EU Better Regulation standards: 3% social discount rate for CBA. Distributional analysis assessing impacts on vulnerable groups. Sensitivity analysis testing robustness of conclusions to assumptions.

ESF+ (European Social Fund Plus) Requirements

Performance Framework: Common output indicators tracked through Outcome Monitoring (Section 3.2.2). Result indicators requiring outcome measurement. Impact evaluation employing Quasi-Experimental Design (Section 3.4.2) minimum, Randomised Controlled Trials (Section 3.4.3) preferred for innovation testing.

Evaluation frequency: Annual progress reports using Foundation Methods (Sections 3.2.1-3.2.3). Mid-term evaluation adding Intermediate Methods (Sections 3.3.1-3.3.3) demonstrating efficiency and social value. Final evaluation including economic evaluation and causal assessment where feasible given programme scale.

4.2 Data Protection and GDPR Compliance

All evaluation activities must comply with General Data Protection Regulation (GDPR) and applicable national data protection laws. GDPR requirements detailed in Chapter 2 apply to all evaluation methods.

Legal basis for processing: Evaluation requires lawful basis under GDPR Article 6. Common bases: Consent (Article 6(1)(a)) for sensitive data, participant interviews, optional collection—requires explicit informed freely-given consent. Legitimate Interest (Article 6(1)(f)) for outcome monitoring using programme data for service improvement—requires a documented balancing test. Public Task (Article 6(1)(e)) for evaluation mandated by public funders or regulatory requirements—requires clear legal mandate (European Union, 2016).

Special category data: Health, ethnicity, religion, sexual orientation (GDPR Article 9) (European Union, 2016) require explicit consent or specific exemptions (scientific research, public health monitoring). Minimise collection. Document necessity.

Data minimisation: Collect only data necessary for specified evaluation purposes. Every data point must answer specific evaluation questions or meet explicit regulatory requirements.

Cross-border data transfers: Multi-country programmes transferring personal data outside the European Economic Area require appropriate safeguards. EU-to-EU transfers follow standard GDPR procedures. EU-to-UK transfers permitted under



adequacy decision. EU-to-third-countries require Standard Contractual Clauses unless the destination country has an EU-approved adequacy decision. Norway, Iceland, Liechtenstein (EEA non-EU members) follow GDPR rules. Switzerland (non-EEA) has separate bilateral agreements but data transfers are treated similarly.

Participant rights: Right to be informed (transparent privacy notice), right of access (request copy of data), right to rectification (correct inaccurate data), right to erasure with limitations for research under Article 89, right to restrict processing, right to object, right to data portability. Programmes must establish procedures responding to rights requests within GDPR timeframes (typically one month).

Resources: Annex A.1 Data Management Plan template meeting HORIZON requirements. Annex A.2 GDPR-compliant consent forms. Annex A.4 Cross-border data transfer protocols including Standard Contractual Clauses.

4.3 Cross-Border Harmonisation for Multi-Country Programmes

Many EU programmes operate across multiple Member States requiring coordinated evaluation enabling meaningful comparison whilst respecting that services operate in different national contexts. Cross-border harmonisation balances standardisation (enabling comparison and aggregation) with local adaptation (respecting context and subsidiarity).

Core challenge: How to compare outcomes across countries when services delivered through different national systems, languages differ, cultural norms affect responses, data collection systems vary, legal frameworks differ in implementation, national teams resist standardisation claiming unique context.

Solution approach: Ex-ante output harmonisation—different questionnaires and procedures designed from start to enable comparison whilst remaining culturally appropriate. Most common approach for HORIZON evaluations: core indicators mandated plus flexibility for adaptation.

Implementation Process

Step 1: Agree core indicators. The multi-country team agrees to the minimum set of indicators all sites must collect identically for cross-country comparison. Typically 3-7 core indicators. Example: "% employed 6 months post-programme" with identical definition, timing, measurement method. Core indicators capture primary outcomes essential for EU-level synthesis.

Step 2: Allow local adaptation. Beyond core indicators, countries adapt methods to context—additional questions, qualitative approaches, culturally appropriate

instruments. Example: Italy adds social enterprise outcomes, Germany tracks apprenticeships, each conducts qualitative research in their own language. Local adaptation enables rich national understanding whilst maintaining core comparability.

Step 3: Translation protocols. Professional translation followed by back-translation followed by cognitive testing ensuring equivalence. Translation is not literal word-for-word but conceptual equivalence ensuring respondents in different languages understand questions identically. Cognitive testing with small samples in each country verifies understanding before full implementation.

Step 4: Coordinated data management. Common database structure, standardised variable names and coding, central quality monitoring. Data coordinator ensures consistency across sites, identifies quality issues early, facilitates cross-country analysis.

Step 5: Multi-level analysis and reporting. Country-specific reports providing contextualised findings. Cross-country synthesis identifying European patterns. Explicit discussion of why differences exist (policy environment, implementation variation, cultural factors, economic context). Both levels of analysis required—neither country reports nor EU aggregate alone sufficient.

Governance and Resources

Governance structure: Central coordination team (overall evaluation management, ensures harmonisation). National evaluation teams (country-level implementation, contextual adaptation). Steering group with country representatives deciding methodological issues (core indicators, harmonisation approach). Regular coordination meetings (quarterly minimum during active data collection).

Budget and timeline additions: Add 30-50% to baseline evaluation costs for coordination, translation, harmonisation. Professional translation and back-translation. Cross-country coordination meetings. Cross-country synthesis analysis. Timeline additions: 3-6 months upfront for harmonisation design, indicator agreement, translation. Quarterly coordination throughout data collection. 2-4 months synthesis phase after country reports.

Common challenges and solutions: National teams resist standardisation claiming unique context—solution: core plus flexibility compromise, "You MUST measure these 5 things identically, CAN measure 10 additional things however you want." GDPR implemented differently across countries—solution: design for most restrictive standards so all can comply, establish data-sharing agreements upfront, use Standard Contractual Clauses for non-EEA transfers. Language barriers—solution: working language (typically English), professional interpretation in key meetings, translate key materials to all languages, use visual tools. Power imbalances where wealthier countries

dominate—solution: explicit governance ensuring all voices heard, rotate responsibilities, equitable budget allocation, capacity-building support.

Quality standards: Harmonisation strategy explicitly documented in written protocol. Core indicators agreed upfront before data collection with all partners. Translation protocols ensuring equivalence documented. Cultural adaptation procedures documented. Regular cross-country coordination documented through meeting notes. Country-specific AND cross-country reports both produced. Context differences explicitly discussed explaining outcome variation. Data protection compliance verified across all countries. Equitable partner participation with all countries contributing meaningfully.

Resources: Annex B.9.1 Core indicator selection framework. Annex B.9.2 Translation protocols. Annex B.9.3 Data harmonisation procedures. Annex B.9.4 Multi-level reporting templates. Annex B.9.5 Cross-border coordination checklist. Annex B.9.6 Cross-border evaluation quality checklist.

4.4 Reporting Standards for EU Programmes

HORIZON deliverable reports: Executive summary (2 pages) providing key findings, implications, recommendations for non-technical audience. Methodology section (5-10 pages) describing methods used referencing relevant Green Book sections, sample characteristics, data collection procedures, limitations. Findings section (15-25 pages) presenting results with supporting evidence (statistics, quotes, case examples) answering evaluation questions directly. Limitations and quality section (3-5 pages) addressing threats to validity, data quality assessment, confidence in findings, alternative explanations considered. Recommendations section practical and evidence-based linked to findings.

ESF+ reporting: Annual progress tracking output indicators and short-term outcomes. Mid-term evaluation demonstrating efficiency and effectiveness. Final evaluation including economic analysis and causal assessment where proportionate.

General principles: Answer evaluation questions specified in grant agreement or call conditions. Reference methodological standards citing relevant Green Book sections demonstrating compliance with established methods. Balance accountability ("Did it work?") and learning ("How can we improve?"). Accessible language maintaining technical rigour with non-technical summaries. Honest about limitations being transparent about what evaluation can and cannot conclude.

4.5 Reference Standards and Resources

EU guidance documents: EU Better Regulation Toolbox

(https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation-toolbox_en). HORIZON Europe Programme Guide. ESF+ Regulation (<https://ec.europa.eu/european-social-fund-plus>). GDPR Official Text (<https://gdpr-info.eu/>). European Commission guidance on cross-border data transfers.

Method-mandate mapping: Theory of intervention requirement met by Theory of Change (Section 3.2.1). Output monitoring requirement met by Outcome Monitoring (Section 3.2.2). Stakeholder consultation requirement met by Stakeholder Feedback (Section 3.2.3). Impact assessment requirement met by Cost-Benefit Analysis (Section 3.4.1) or Multi-Criteria Decision Analysis (Section 3.3.2). Causality evidence requirement met by Quasi-Experimental Design (Section 3.4.2) or Randomised Controlled Trials (Section 3.4.3). Value for money requirement met by Cost-Effectiveness Analysis (Section 3.3.1) or Cost-Benefit Analysis (Section 3.4.1). Transferability understanding met by Realist Evaluation (Section 3.4.4). Cross-border coordination met by protocols in this chapter and Annexes B.9.1-B.9.6.

Chapter 5:

Presentation of results



CHAPTER 5: PRESENTATION OF RESULTS

5.1 Purpose of Results Presentation

Results presentation provides objective evidence and analysis feeding into design, scrutiny, and approval processes supporting programme and policy decisions. Appraisal and evaluation results should be presented transparently showing clearly the social value of interventions and enabling informed decisions about continuation, adaptation, or scaling.

Results' presentation serves multiple audiences requiring different levels of technical detail. Senior decision-makers require concise executive summaries highlighting key findings and recommendations. Technical specialists require detailed methodological documentation enabling assessment of analytical quality. Funders require evidence demonstrating compliance with requirements and value for money. Stakeholders including participants, staff, and partners require accessible summaries demonstrating how findings inform improvements.

Beyond these internal audiences, there is a broader case for making evaluation findings accessible to the general public. Transparent presentation of results strengthens democratic accountability, builds public support for effective programmes, and contributes to informed public discourse about social service investment. Where appropriate, programmes should produce publicly available summaries using accessible language and visual presentation.

5.2 Core Principles for Results Presentation

Transparency: Present results clearly showing assumptions, data sources, analytical choices, and limitations. Cross-reference summary statements to detailed evidence in the main body of the report. Key data and assumptions should be identified and cross-referenced to original sources. Decision-makers should be able to trace how conclusions were reached.

Accessibility: Balance technical rigour with accessible communication. Provide executive summaries (2-3 pages) for non-technical audiences summarising key findings, implications, and recommendations without jargon. Full reports should include technical detail enabling specialist assessment whilst remaining readable for informed non-specialist audiences.

Completeness: Report positive findings, null findings, and negative findings with equal prominence. Absence of expected effects or evidence of harm requires transparent reporting not defensive dismissal. Present both monetised benefits and non-

monetisable outcomes. Acknowledge uncertainty explicitly through confidence intervals, sensitivity analysis results, and discussion of limitations.

Actionability: Connect findings to practical decisions. Recommendations should be specific, evidence-based, and address questions decision-makers face. Generic recommendations ("continue monitoring outcomes") provide less value than specific guidance ("modify intake criteria to prioritise participants with baseline characteristic X, which predicts 40% higher success rates").

Proportionality: Reporting effort should match programme scale and decision stakes. Small programmes require brief reports (10-20 pages) focusing on essential findings. Large programmes or high-stakes decisions justify comprehensive reports (40-80 pages) with detailed technical documentation. Excessive documentation for small decisions wastes resources; inadequate documentation for major decisions fails accountability standards.

5.3 Report Structure

Executive Summary (2-3 pages)

Purpose: Enable senior decision-makers to understand key findings and implications within 10 minutes reading time.

Essential elements: Programme context and evaluation questions (1 paragraph). Methods used briefly (1-2 sentences referencing detailed methods section). Key findings directly answering evaluation questions (3-5 bullet points with supporting evidence). Practical implications for programme continuation, adaptation, or scaling. Specific recommendations with implementation guidance. Significant residual risks or uncertainties affecting recommendations.

What to avoid: Technical jargon without definition. Detailed methodology. Raw data tables. Hedging language obscuring conclusions ("findings suggest possible indication that outcomes may have improved"). Long narrative paragraphs—use structured headings and brief bullet points.

Main Body

1. Introduction (3-5 pages)

Programme description: target population, services provided, implementation model, geographic scope, duration, funding. Strategic context explaining programme's fit within broader policy objectives or organisational strategy. Evaluation questions specifying what evaluation seeks to answer. Constraints and dependencies where relevant.

2. Methods (5-10 pages for Foundation/Intermediate Methods, 10-15 pages for Advanced Methods)

Evaluation design with clear justification referencing relevant Green Book sections. Data sources describing recruitment, sampling, instruments used. Data collection procedures including timing, administration protocols, response rates. Analytical approach explaining how data is analysed to answer evaluation questions. Quality assurance procedures. Limitations explicitly acknowledged including threats to validity, incomplete data, constraints on causal inference.

Reference relevant Green Book sections demonstrating compliance with established standards. For commissioned specialist evaluations, include specialist credentials and independence statements.

3. Findings (15-30 pages depending on complexity)

Present results systematically addressing each evaluation question. Support findings with appropriate evidence: quantitative results with confidence intervals where applicable, qualitative data with illustrative quotes, case examples demonstrating mechanisms or patterns, visual presentation (charts, graphs, tables) complementing text.

Distinguish clearly between descriptive findings (what happened), associational findings (what correlates with outcomes), and causal findings (what programme caused). Avoid causal claims unless employing methods establishing causation (Sections 3.4.2-3.4.3).

Present disaggregated analysis where relevant: outcomes by participant characteristics, implementation variation across sites, differential effects in different contexts. Disaggregation reveals for whom programmes work best and where adaptations are needed.

4. Limitations and Quality (3-5 pages)

Threats to validity: selection bias, measurement error, confounding factors, attrition, contextual changes affecting interpretation. Data quality assessment: completeness rates, consistency checks, outliers, missing data handling. Confidence in findings: which conclusions are well-supported, which tentatively require additional evidence. Alternative explanations: rival hypotheses that could explain observed patterns, evidence for and against alternatives.

Limitations discussion should be honest and specific, not perfunctory disclaimers. Decision-makers benefit from understanding what evaluation can and cannot conclude with confidence.

5. Conclusions and Recommendations (5-10 pages)

Direct answers to evaluation questions stated explicitly. Practical implications: what findings mean for programme continuation, adaptation, scaling, or discontinuation. Specific recommendations: evidence-based, actionable, addressing decisions programme faces. Implementation guidance: how to act on recommendations including priority sequence, resource requirements, potential barriers. Learning for future evaluations: what worked well methodologically, what would improve future efforts.

Recommendations should follow logically from findings. Each recommendation should reference specific evidence supporting it. Avoid generic recommendations disconnected from actual findings.

Technical Annexes

Detailed technical documentation supporting the main body: complete data collection instruments, sample recruitment protocols and response rates, detailed statistical analyses including sensitivity tests, cost accounting details showing all assumptions, Theory of Change visual and narrative, stakeholder consultation records. Technical annexes enable specialist reviewers to assess analytical quality whilst keeping the main body readable.

5.4 Visual Presentation

Tables: Present summary statistics, comparisons across alternatives, disaggregated results. Tables should be self-explanatory with clear titles, column/row labels, units specified, notes explaining abbreviations or special coding. Complex tables in annexes; simplified versions in main body.

Charts and graphs: Bar charts for comparisons, line graphs for trends over time, scatter plots for relationships, confidence intervals showing uncertainty. Visual presentation should clarify not merely illustrate—poorly designed visuals add confusion not clarity. All charts require clear titles, axis labels, legends, and source notes.

Appraisal Summary Tables: For economic evaluations (CEA, CBA, MCDA), summary tables present key metrics enabling rapid comparison of alternatives. Tables should show: Net Present Value or cost-effectiveness ratios for each alternative, confidence intervals showing uncertainty, sensitivity analysis results, non-monetised benefits, significant residual risks. Present figures in absolute terms not merely as increments from business-as-usual, enabling transparent comparison. See UK Treasury Green Book Annex for template examples.

5.5 Presenting Uncertainty and Sensitivity Analysis

Confidence intervals: Where statistical analysis permits, report confidence intervals alongside point estimates. Example: "Cost per outcome £4,200 (95% CI: £3,400-£5,300)" more informative than point estimate alone. Wide confidence intervals signal substantial uncertainty requiring cautious interpretation.

Sensitivity analysis results: Test how conclusions change when varying key assumptions. Present results showing: which parameters affect conclusions materially, threshold values where conclusions reverse ("If completion rate falls below 65%, programme becomes more costly per outcome than alternative"), robust conclusions (findings hold across plausible parameter ranges). Sensitivity analysis distinguishes confident conclusions from tentative findings requiring additional evidence.

Scenario analysis: For complex evaluations with multiple uncertain parameters, present optimistic, expected, and pessimistic scenarios. Scenario analysis helps decision-makers understand best case, most likely case, and worst-case possibilities.

Transparent discussion: Acknowledge what evaluation cannot conclude with confidence. Decision-makers benefit from understanding which findings are well-established, which findings tentative, and what additional evidence would strengthen conclusions. Honesty about uncertainty builds credibility; false precision undermines trust.

5.6 Reporting Non-Monetised Outcomes

Many social service outcomes resist monetary valuation or stakeholders object to monetisation. Dignity, justice, voice, community cohesion, cultural preservation represent genuine value not captured in cost-benefit ratios. Present non-monetised outcomes explicitly not hidden.

Qualitative description: Describe outcomes resisting quantification using rich qualitative evidence. Example: participant testimonials demonstrating restored dignity, staff observations of increased community engagement, case studies showing life transformation.

Structured assessment: Where outcomes can be assessed systematically even if not monetised, present structured evidence. Example: rubrics assessing quality of life dimensions, validated scales measuring wellbeing without monetary conversion, systematic content analysis of participant narratives.

Relative importance: Discuss importance of non-monetised outcomes relative to monetised benefits. Decision-makers should understand full value created including



dimensions resisting monetary valuation. Cost-benefit ratio provides a partial not complete picture of social value.

5.7 Presenting Distributional Effects

Who benefits: Disaggregate outcomes by participant characteristics (age, gender, baseline severity, socioeconomic status, ethnicity where relevant and measured). Assess whether benefits accrue primarily to initially advantaged or disadvantaged groups. Progressive interventions disproportionately benefit disadvantaged; regressive interventions disproportionately benefit advantaged. Both patterns may be justified depending on programme objectives, but transparency enables informed decisions.

Who pays: Identify who bears costs. Public sector programmes funded by taxpayers. Participant time contributions represent real costs even when unpaid. Opportunity costs affect other programmes competing for the same resources. Transparent cost distribution enables assessment of fairness alongside efficiency.

Geographic distribution: For place-based programmes, present outcomes and costs by area. Some regions may benefit more than others due to implementation variation, population characteristics, or contextual factors. Geographic distribution analysis informs scaling and adaptation decisions.

Equity implications: Assess whether the programme reduces, maintains, or increases inequalities. Interventions may be cost-effective on average whilst exacerbating disparities if benefits accrue primarily to advantaged groups. Equity analysis alongside efficiency analysis provides a comprehensive basis for policy decisions.

5.8 Reporting for Different Audiences

Senior decision-makers: Executive summary (2-3 pages maximum). Key findings as brief bullet points. Clear recommendation with supporting rationale. Visual presentation (single chart or table showing core comparison). Practical next steps.

Technical specialists: Complete methodology section. Detailed analytical procedures. Sensitivity analysis results. Quality assessment discussion. Technical annexes with complete documentation.

Funders: Compliance evidence demonstrating adherence to funder requirements. Value for money analysis. Outcome indicators specified in grant agreements. Budget versus actual expenditure. Recommendations for programme continuation or adaptation.

Stakeholders (participants, staff, partners): Accessible summary (3-5 pages). Plain language avoiding jargon. Visual presentation (infographics, photos, charts). Emphasis

on how findings inform improvements. "You said, we did" format showing how stakeholder feedback influenced conclusions. Opportunities for dialogue about findings and recommendations.

Academic audiences: Comprehensive methodology. Theoretical framing. Literature review situating findings in broader evidence base. Detailed analytical procedures enabling replication. Limitations discussion with attention to methodological debates. Contribution to knowledge claims.

Different audiences require different documents. Producing a single report attempting to serve all audiences simultaneously serves none well. Develop targeted outputs for each primary audience.

5.9 Quality Standards for Reports

Completeness: Report addresses all evaluation questions specified at outset. Presents positive findings, null findings, negative findings with equal prominence. Includes executive summary, methods, findings, limitations, conclusions. Provides sufficient detail enabling assessment of analytical quality.

Accuracy: Data reported accurately with sources cited. Calculations verified by an independent reviewer. Claims supported by presented evidence. Quotations and case examples accurately represent underlying data.

Clarity: Writing accessible to intended audiences. Technical terms defined. Acronyms spelled out at first use. Visual presentation enhances rather than confuses understanding. Logical flow from evaluation questions through methods to findings to conclusions.

Balance: Presents evidence for and against conclusions. Acknowledges alternative explanations. Discusses strengths and limitations with equal attention. Avoids advocacy for particular conclusions unsupported by evidence.

Utility: Findings inform actual decisions. Recommendations actionable and evidence-based. Report delivered when decision-makers need it—late reports miss decision windows. Format and length proportionate to stakes.

Reproducibility: Evaluation methodology and anonymised data should be documented with sufficient detail to enable independent replication of analyses. Where feasible and consistent with data protection requirements, anonymised datasets and analytical code should be made available to support verification and secondary research.

5.10 Common Reporting Errors

Error 1: Burying key findings. Essential conclusions hidden in the middle of dense text. Solution: Executive summary states key findings clearly. Main findings section starts with direct answers to evaluation questions.

Error 2: Excessive hedging. Every statement qualified until meaning obscured ("tentative findings suggest possible indication that outcomes may have potentially improved"). Solution: State conclusions clearly with appropriate confidence level. Distinguish well-supported conclusions from tentative findings.

Error 3: Disconnected recommendations. Recommendations generic or not traceable to specific findings. Solution: Each recommendation explicitly references evidence supporting it. Recommendations follow logically from analysis.

Error 4: Missing limitations discussion. The report presents findings without acknowledging threats to validity or uncertainties. Solution: Explicit limitations section discussing what evaluation can and cannot conclude confidently.

Error 5: Inappropriate length. Small programmes receive an 80-page report; major policy decisions receive a 15-page summary. Solution: Match report length and technical depth to programme scale and decision stakes.

CONCLUSIONS

This European Green Book for Social Services establishes a proportionate, standards-based evaluation framework addressing the distinctive challenges of social service contexts within the European Union. Three principal conclusions emerge from the framework's development.

First, proportionate evaluation is achievable at every scale. The three-tier architecture demonstrates that credible evaluation does not require resources beyond the reach of small and medium-sized social service providers. Foundation Methods — Theory of Change, Outcome Monitoring, and Stakeholder Feedback — provide a rigorous evidence base using internal capacity and modest external support. The critical barrier is not methodological complexity but systematic implementation: establishing baseline measurement before programme delivery, maintaining consistent data collection, and using findings to inform decisions. Programmes implementing Foundation Methods will produce evidence that is substantially more valuable than programmes attempting Advanced Methods with insufficient resources.

Second, evaluation serves both accountability and learning. This framework positions evaluation not merely as a compliance requirement for funders but as a mechanism for programme improvement and democratic accountability. The feedback loops embedded in Foundation Methods — particularly the stakeholder engagement cycle of collection, analysis, adaptation, and communication — constitute adaptive management in practice. Programmes that evaluate systematically are better positioned to demonstrate value, defend funding, adapt to changing contexts, and improve outcomes for the people they serve.

Third, the framework must be tested, adapted, and refined through implementation. The standards and methods presented here draw on established evaluation methodology adapted for social service contexts, but their practical utility will be demonstrated through application. The BENEFITS project's pilot testing (D2.3) across seven EU Member States will provide empirical evidence on implementation feasibility, identify barriers specific to different national and organisational contexts, and inform revisions ensuring the framework serves its intended users. The Green Book is designed as a living document: its value lies not in theoretical comprehensiveness but in practical applicability for organisations seeking to evaluate their work credibly and proportionately.

The evaluation gap in European social services is not primarily a methodological problem — robust methods exist and are presented here. It is an implementation problem: building capacity, establishing data systems, and embedding evaluation into organisational culture. This Green Book provides the methodological foundation. Closing the gap requires sustained investment in evaluation capacity across the sector, supported by funders who resource evaluation proportionately and commissioners who value evidence alongside service delivery.

Case Studies



CASE STUDIES

CASE STUDY 1: Theory of Change

Table 8 - Theory of Change Programme Snapshot

Element	Description
Programme	Ohjaamo One-Stop Youth Guidance Centres
Location	Finland (national programme, ~70 centres as of 2022)
Scale	Large – approximately 140,000 young participants in 2019; ~1,000 professionals
Duration	2014-present (10 years operational)
Target Population	Young people under the age of 30 not in employment, education, or training (NEET), or at risk of becoming NEET
Key Innovation	Multi-agency coordination in single physical location; cross-sectoral services under one roof
Annual Budget	Helsinki centre: €3.1 million (75% ESF, 25% City of Helsinki). National government funding: €5 million per year (post-2018). Additional ESF project funding across network
Funding	European Social Fund (Youth Employment Initiative), Finnish Ministry of Education and Culture, Finnish Ministry of Economic Affairs and Employment, municipalities

THE CHALLENGE

Finland's Public Employment Services (PES) offered extensive support for young people, but studies commissioned by the Ministry of Economic Affairs and Employment revealed that young people preferred face-to-face services over the online support offered by PES. Services were fragmented, with limited coordination between providers. Young people with multiple support needs had to make separate appointments with different offices, meaning no single provider held a holistic overview of their situation.

Youth unemployment and NEET rates remained a persistent concern. The existing system was organised around institutional boundaries rather than the needs of young people, creating gaps that were particularly problematic for those facing multiple barriers to education and employment.

THE INTERVENTION

Ohjaamo centres integrate cross-sectoral services in a single accessible location. Staff from different organisations — including PES youth counsellors, social workers, nurses, psychologists, outreach workers, and study counsellors — provide services together under one roof.

Core model features included multi-agency co-location, with a single physical space housing staff from multiple partner agencies across public, private, and third-sector organisations. Services were provided through multiple channels: face-to-face, electronic, online, and telephone. Young people could visit centres without an appointment.

The environment was designed to be informal, non-discriminatory, and welcoming. Young people participated in developing centre operations, including being interviewed about service design. Peer support and outreach youth work were typical features.

Services varied locally — each centre adapted to geographic location and municipal conditions. The organisational model differed across centres, with the definition deliberately left flexible during the development phase. Service domains included employment counselling, social work, health services, educational guidance, recruitment events, and support for social skills and daily living.

WHY THEORY OF CHANGE WAS SELECTED

Theory of Change was an appropriate evaluation framework for Ohjaamo because the programme operates through multiple interconnected mechanisms — reduced fragmentation, holistic assessment, coordinated case management, and reduced stigma — requiring explicit articulation before outcomes can be meaningfully measured. The OECD's *Investing in Youth: Finland* report emphasised that good evaluations are critical to promote evidence-based policy-making, and the programme's complex multi-agency model required understanding of causal pathways, not merely outcome tracking.

The INNOSI (Innovative Social Investment) research project, funded under the EU's Horizon 2020 programme, developed a Theory of Change for Ohjaamo as part of a comparative case study of social investment innovations across Europe. The ToC development was documented in the INNOSI Work Package 4 case study report, providing a detailed account of how programme logic, assumptions, and intermediate outcomes were articulated for the Finnish Youth Guarantee context.

HOW THE THEORY OF CHANGE DEVELOPMENT WAS CONDUCTED

Evaluators: INNOSI consortium researchers (Horizon 2020 project), working with the Association of Finnish Local and Regional Authorities and the ESF-funded Kohtaamo coordination project.



Existing programme logic: The INNOSI researchers began by documenting the existing theory of change implicit in Ohjaamo's design. This included the core assumptions underlying the model: that co-location reduces barriers to access; that multi-agency working produces better outcomes than siloed services; that young people respond better to informal, low-threshold environments; and that holistic support addressing multiple life domains simultaneously is more effective than sequential referrals.

New Theory of Change development: The INNOSI team developed a formal ToC framework articulating the programme's inputs (ESF funding, municipal resources, multi-agency staff), activities (guidance, counselling, workshops, outreach), outputs (service contacts, referrals, action plans), intermediate outcomes (improved service coordination, increased young people's agency, reduced time to access support), and long-term outcomes (employment, education entry, social inclusion, reduced NEET rates).

Assumptions and justifications: The ToC process identified critical assumptions requiring testing, including that co-location would lead to genuine collaboration rather than merely co-existence; that local flexibility would produce innovation rather than inconsistency; and that young people would use the services if made accessible. These assumptions were documented alongside the evidence base supporting them and the risks if they proved incorrect.

Participatory approach: The evaluation conducted during 2018–2020 was participatory and development-oriented. The evaluation summarised main observations and made recommendations for developing and consolidating the support structure for guidance centres. Academic researchers examined Ohjaamo through multiple lenses, including observational research examining guidance situations (n=68) using cluster analysis to identify distinct guidance practices.

KEY FINDINGS

Programme uptake and scaling: From initial pilot centres opened in 2015, the network expanded to approximately 70 centres across Finland employing around 1,000 professionals. In 2019, the network served approximately 140,000 young participants.

Government institutionalisation: The Finnish Government decided to make Ohjaamo a permanent practice, providing €5 million per year in dedicated funding. The programme was placed on a permanent footing as a Youth Guarantee implementation tool, representing a significant signal of policy confidence.

EU recognition: The European Commission identified Ohjaamo as good practice for Youth Guarantee implementation. Commissioner Nicolas Schmit visited Helsinki's Ohjaamo centre in October 2022, highlighting the model's effectiveness and ESF contribution.

OECD assessment: The OECD's *Investing in Youth: Finland* report recognised Ohjaamo as a significant innovation but noted that good evaluations are critical for evidence-based policy-making and that Finland could learn from countries that include requirements for programme performance tracking and impact evaluation in funding legislation.

Research evidence: Academic research has described Ohjaamo centres as "the most ambitious investment on the national level in the provision of multi-agency youth services in Finland." However, empirical and theoretical research evidence on multi-agency guidance remains limited, as noted by Finnish researchers. The ToC development through INNOSI provided one of the first formal articulations of the programme's causal logic.

KEY DATA SUMMARY

Indicator	Value
Centres operational	~70 (as of 2022)
Young participants (2019)	~140,000
Professionals employed	~1,000
Helsinki centre ESF funding	€3.1 million (75% ESF)
National government funding	€5 million/year

SOURCES

- European Social Fund Plus (n.d.). One-stop-shop guidance centres for young people (Ohjaamo). European Commission. <https://european-social-fund-plus.ec.europa.eu/en/social-innovation-match/case-study/one-stop-shop-guidance-centres-young-people-ohjaamo>
- INNOSI (2017). *Work Package 4 Case Study Report: Finland Youth Guarantee / Ohjaamo*. Horizon 2020 project. Association of Finnish Local and Regional Authorities.
- Kautto, T., Korpilauri, T., Pudas, M. and Savonmäki, P. (2018). *One-Stop Guidance Center (Ohjaamo)*. Ed. Määttä, M. Available at: <http://www.doria.fi/handle/10024/162148>
- OECD (2019). *Investing in Youth: Finland*. OECD Publishing, Paris.
- Määttä, M. (2019). 'Reforming youth transition support with the multi-agency approach? A case study of the Finnish one-stop guidance centers.' *Sociologija*, 61(2), 277–291. <https://doi.org/10.2298/SOC1902277M>

CASE STUDY 2: Cost-Effectiveness Analysis

Table 9 - Cost Effectiveness Analysis Programme Snapshot

Element	Description
Programme	Individual Placement and Support (IPS) — supported employment for people with severe mental illness
Location	Three locations in Denmark (Copenhagen area and Southern Denmark)
Scale	Large — 720 participants in a three-arm randomised clinical trial
Duration	2012–2018 (recruitment and follow-up); IPS delivery ongoing in Danish mental health services
Target Population	People diagnosed with severe mental illness (schizophrenia spectrum disorders, bipolar disorder, recurrent depression)
Funding	Danish Agency for Labour Market and Recruitment, TrygFonden (Danish foundation), Obel Family Foundation
Key Innovation	Integration of employment specialists directly within mental health treatment teams; rapid job search in the competitive labour market rather than pre-vocational training
Trial Registration	ClinicalTrials.gov NCT01722344

THE CHALLENGE

People with severe mental illness (SMI) face among the highest unemployment rates of any population group. Employment is consistently identified as a priority by people living with SMI, and research demonstrates that work is associated with improved mental health, social inclusion, and recovery. Yet traditional vocational rehabilitation — typically involving extended pre-vocational assessment, sheltered work, and gradual progression toward competitive employment — has shown limited effectiveness in helping people with SMI enter and sustain competitive jobs.

At baseline, participants in the Danish trial had a mean age of 32.8 years, and 39% had completed only elementary or lower secondary school. The majority (76.5%) had a schizophrenia spectrum disorder, with the remainder diagnosed with bipolar disorder (12%) or recurrent depression (11%). These are conditions associated with substantial healthcare costs, high public transfer payments, and significant income loss.

THE INTERVENTION

Individual Placement and Support (IPS) follows an evidence-based model with defined principles: competitive employment as the primary goal; eligibility based on client choice (no exclusion criteria based on diagnosis or symptoms); rapid job search in the open labour market rather than lengthy pre-vocational assessment; integration of employment specialists within mental health treatment teams; attention to client preferences regarding job type, hours, and location; and time-unlimited, individualised support after job placement.

The Danish trial compared three conditions: (a) IPS alone (n=243); (b) IPS supplemented with cognitive remediation and work-focused social skills training, termed IPSE (n=238); and (c) service as usual, SAU (n=239). Participants allocated to SAU continued to receive counselling at municipal job centres and treatment in early intervention teams (OPUS teams) or community mental health treatment teams.

WHY COST-EFFECTIVENESS ANALYSIS WAS CHOSEN

Policymakers and administrators were increasingly interested in IPS as a route to improving employment outcomes for people with SMI, but needed evidence on whether the approach represented good value for money — particularly given that IPS requires dedicated employment specialists integrated within clinical teams. A cost-effectiveness analysis was essential to demonstrate not only whether IPS produced better outcomes but also whether the additional investment was justified relative to existing services.

The trial's design — a large-scale, three-arm randomised clinical trial with 720 participants — provided the ideal platform for a rigorous economic evaluation. The availability of Danish national register data covering healthcare costs, municipal social care costs, and labour market service costs enabled a comprehensive societal perspective without relying on self-reported resource use.

HOW THE COST-EFFECTIVENESS ANALYSIS WAS CONDUCTED

Evaluators: Thomas Nordahl Christensen, Marie Kruse (Danish Centre for Health Economics, University of Southern Denmark), Lone Hellström, and Lene Falgaard Epløv (Copenhagen Research Centre for Mental Health — CORE). Published in *European Psychiatry* (2020, vol. 64, e3).

Study design: Economic evaluation alongside a pre-registered, three-arm, superiority randomised clinical trial (NCT01722344).

Cost data: Extracted from nationwide Danish registers covering the full 18-month follow-up period. Cost categories included: healthcare costs (psychiatric hospital care, somatic hospital care, primary care, prescription medication); municipal social care costs (home care, residential care, social services); labour market service costs (job

centre contacts, activation programmes, sheltered employment); and IPS/IPSE intervention costs (employment specialist time, cognitive remediation sessions, social skills training).

Effectiveness measures — dual outcomes:

Cost-utility analysis: Quality-adjusted life years (QALYs) derived from participants' responses to the EuroQol five-dimensional questionnaire (EQ-5D) at baseline and 18-month follow-up. The EQ-5D measures five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Danish population tariffs were applied to generate utility scores.

Cost-effectiveness analysis: Hours in competitive employment or education during the 18-month follow-up, derived from national registers (100% follow-up rate for register-based outcomes).

Statistical approach: Incremental cost-effectiveness ratios (ICERs) were computed for both QALYs and hours in employment. Bootstrapping (nonparametric) was used to handle sampling uncertainty. Results were presented on cost-effectiveness planes and as cost-effectiveness acceptability curves. Complete case analysis was the primary approach (n=462 with EQ-5D data at both time points, 64% of total sample); sensitivity analysis was conducted using multiple imputation.

KEY FINDINGS

Three trial arms: IPS = Individual Placement and Support; IPSE = IPS supplemented with cognitive remediation and social skills training; SAU = Service as Usual (standard municipal job centre counselling plus mental health treatment).

Cost findings (18-month follow-up, per participant):

Cost category	IPS	IPSE	SAU
Psychiatric hospital care	€3,730 lower than SAU	€4,545 lower than SAU	Reference
Total costs (societal perspective)	€9,543 lower than SAU (significant)	€7,288 lower than SAU (significant)	Reference

Both IPS and IPSE generated statistically significant cost savings compared to service as usual. The savings were driven primarily by reduced psychiatric hospital care — participants receiving IPS spent substantially less time in psychiatric inpatient settings than those in the control group.

Effectiveness findings (from main trial, JAMA Psychiatry 2019):

During the 18-month follow-up, the IPSE group worked or studied a mean of 488.1 hours (SD 735.6), compared with 340.8 hours (SD 573.8) in the SAU group (success-rate difference 0.151, 95% CI 0.01–0.295, $p=0.016$). The IPS group worked or studied a mean of 411.0 hours (SD 656.9) (success-rate difference 0.127, 95% CI –0.017 to 0.276, $p=0.004$). There was no difference between IPS and IPSE in any vocational outcomes. No differences were found in non-vocational outcomes (symptoms, social functioning, self-esteem) except that both IPS groups reported significantly higher satisfaction with services.

Cost-effectiveness findings:

ICERs did not reach statistical significance, but there was a tendency toward both IPS and IPSE being *dominant* — that is, producing better outcomes at lower cost compared to service as usual. On the cost-effectiveness plane, the majority of bootstrapped cost-effect pairs fell in the south-east quadrant (lower costs, higher effects), indicating that IPS tends to be cost-saving while generating health and employment benefits.

The 30-month follow-up (Christensen et al., 2023) confirmed that register-based employment and education outcomes were sustained beyond the 18-month intervention period, with 100% follow-up achieved through national registers.

KEY DATA SUMMARY

Indicator	Value
Participants randomised	720 (IPS: 243; IPSE: 238; SAU: 239)
Mean age at baseline	32.8 years (SD 9.9)
Sex	38.3% women
Diagnosis: schizophrenia spectrum	76.5%
Follow-up (register data)	100% at 18 and 30 months
Follow-up (EQ-5D)	64% (462 of 720)
Cost saving: IPS vs SAU	€9,543 (significant)
Cost saving: IPSE vs SAU	€7,288 (significant)
Psychiatric hospital saving: IPS vs SAU	€3,730 per person
Psychiatric hospital saving: IPSE vs SAU	€4,545 per person
Hours worked/studied (IPSE vs SAU)	488.1 vs 340.8 ($p=0.016$)
ICER trend	Dominant (lower cost, higher effect)

SOURCES

- Christensen, T.N., Kruse, M., Hellström, L. and Eplov, L.F. (2020). 'Cost-utility and cost-effectiveness of individual placement support and cognitive remediation in people with severe mental illness: Results from a randomized clinical trial.' *European Psychiatry*, 64(1), e3. <https://doi.org/10.1192/j.eurpsy.2020.111>
- Christensen, T.N., Wallstrøm, I.G., Stenager, E., Bojesen, A.B., Gluud, C., Nordentoft, M. and Eplov, L.F. (2019). 'Effects of Individual Placement and Support Supplemented With Cognitive Remediation and Work-Focused Social Skills Training for People With Severe Mental Illness: A Randomized Clinical Trial.' *JAMA Psychiatry*, 76(12), pp.1232–1240. <https://doi.org/10.1001/jamapsychiatry.2019.2291>
- Christensen, T.N., Wallstrøm, I.G., Stenager, E., Hellström, L., Bojesen, A.B., Nordentoft, M. and Eplov, L.F. (2023). '30-Month Follow-Up of Individual Placement and Support (IPS) and Cognitive Remediation for People with Severe Mental Illness: Results from a Randomized Clinical Trial.' *Psychiatry Journal*, 2023, 2789891.
- Drake, R.E., Bond, G.R. and Becker, D.R. (2012). *Individual Placement and Support: An Evidence-Based Approach to Supported Employment*. New York: Oxford University Press.

CASE STUDY 3: Cost-Benefit Analysis

Table 10 - Cost Benefit Analysis Programme Snapshot

Element	Description
Programme	Housing First Pilots (England)
Location	Greater Manchester, Liverpool City Region, West Midlands combined authority areas
Scale	1,387 people received support; 1,061 provided with housing
Duration	2018/19–2023/24 (pilot delivery, with phased start-up across regions); extended into 2024/25 through another funding. Evaluation data to December 2022
Target Population	Homeless people with multiple and complex needs, typically co-occurring mental health issues and substance use
Total Investment	£28 million initial (Autumn 2017 Budget) + £13.9 million extension (from 2022). Total expenditure by December 2022: £27.6 million
Key Innovation	Immediate access to independent, permanent housing with flexible, intensive support—no treatment or sobriety prerequisites; Housing First principles (client choice, harm reduction, recovery orientation)
Report	<i>Evaluation of the Housing First Pilots: Cost Benefit Analysis — Final Report, October 2024</i>

THE CHALLENGE

People experiencing chronic homelessness with multiple and complex needs are among the most disadvantaged populations in England. Traditional "staircase" or "treatment first" approaches require people to become "housing ready" before offering permanent accommodation — moving through outreach, congregate supported housing, then resettlement. This approach often fails for people with the most complex needs.

Client profile data from baseline surveys (n=312) revealed the severity of disadvantage: 96% had experienced rough sleeping; 48% had not been in settled housing for five or more years; 80% self-reported depression and 79% anxiety; 71% had used drugs in the previous three months; 75% had spent time in prison; and only 57% were registered with a GP.

THE INTERVENTION

Housing First provides immediate, independent, permanent housing with personalised, flexible, non-time-limited support. No preconditions around "housing readiness" or participation in treatment are imposed. The approach is based on seven principles: right to a home; flexible support for as long as needed; housing and support separated; individual choice and control; active engagement; strengths-based practice; and harm reduction.

Three Pilot models were established, each with different delivery structures: Greater Manchester (consortium-led, delivered by Great Places Housing Group under contract to GMCA; budget £8.0 million); Liverpool City Region (delivered in-house by Combined Authority with directly recruited staff; budget £7.7 million); and West Midlands (locally commissioned across 7 metropolitan borough councils, mix of in-house and contracted delivery; budget £9.6 million).

WHY COST-BENEFIT ANALYSIS WAS CHOSEN

The Pilots represented a £28 million public investment requiring evidence of value for money. CBA was appropriate because Housing First involves quantifiable costs (programme delivery, staffing, housing) and benefits (reduced public service use, improved wellbeing); the target population generates high costs across multiple public services (homelessness, health, criminal justice); and decision-makers needed a single metric (benefit-cost ratio) to assess whether the approach warranted further investment. The CBA framework was agreed with MHCLG in December 2019 and shared with the Pilots before data collection commenced.

HOW THE COST-BENEFIT ANALYSIS WAS CONDUCTED

Evaluator: ICF Consulting-led consortium commissioned by MHCLG. Dedicated CBA report published October 2024.

Cost assessment: Cost data were provided directly by the three Pilots and analysed by ICF. The framework distinguished between: core staffing costs (salaries, pensions, NI of Housing First support workers); wider staffing costs (governance, administration); overheads (IT, office, training); procured services (mental health, psychology, 24/7 support); personalisation funds (discretionary participant expenses); and in-kind costs (volunteer time, partner staff time for governance). Housing costs (rents, furnishing, refurbishment) were tracked separately. Unit costs were calculated by dividing total costs by both (a) number of clients housed and (b) number of clients receiving any support.

Benefit assessment — two approaches:

Expected benefits: Based on published evidence of the costs of homelessness (Pleace and Culhane, 2016; CSJ, 2021), estimating that providing secure housing for a previously homeless person yields annual public service cost reductions of £10,900–£15,900 (central value £13,400 at 2022 prices). Expected wellbeing benefits valued at £13,289 per person using published evidence on the value of alleviating homelessness.

Observed benefits to date: Based on surveys of Housing First clients at baseline, 6-month, and 12-month follow-up (n=312 at baseline; n=167 at 12 months). Changes in public service use quantified using published unit costs. Wellbeing improvements measured using the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS) and valued at £6,246 per person over 12 months.

Adjustments: Benefits were reduced by 15% to account for clients losing contact or experiencing negative outcomes (based on 3-year data showing deaths, imprisonment, and loss of contact). A further 30% reduction was applied for non-additional outcomes (deadweight), based on international evaluation evidence estimating that 30% of recipients would have achieved secure housing without the programme. The Pilots themselves argue this overestimates deadweight given the severity of their cohort's needs.

Limitations acknowledged: No comparison group was available. Costs are known with high certainty; benefits are subject to significant data gaps and assumptions. The CBA is explicitly described as indicative.

KEY FINDINGS

Total expenditure: £27.6 million by 31 December 2022, with a further £0.4 million in-kind costs. More than 80% spent on staffing, principally Housing First support workers.

Unit costs per person supported per year (annual average, 2018/19–2022/23):

Pilot	Annual cost per person supported	Annual cost per person housed
Greater Manchester	£7,862	£9,524
Liverpool City Region	£12,613	£22,027
West Midlands	£5,558	£7,116
All Pilots average	£7,737	£10,915

The wide variation in unit costs reflects differences in starting points, delivery models (in-house vs. commissioned), set-up costs (Liverpool built delivery infrastructure from scratch, including recruiting a psychology service and establishing multi-disciplinary panels), and COVID-19 disruption.

Cumulative cost per person housed: £26,348 average (range: £18,430 in West Midlands to £43,841 in Liverpool).

Benefits (adjusted for 15% negative outcomes and 30% deadweight):

Benefit type	Expected (annual, per person)	Observed at 12 months (per person)
Reduced public service costs	£7,973	£4,712
Wellbeing benefits	£7,907	£3,716
Total	£15,880	£8,429

Observed public service savings at 12 months comprised homelessness service cost reductions of £6,116 per person per year and prison cost reductions of £1,804 per person per year. No significant change in physical or mental health service use was observed at this stage.

Benefit-cost ratio: 2.1:1 (based on expected benefits); **1.1:1** (based on 12-month observed benefits). Housing Benefits excluded as transfer payments. The evaluation concludes that the Pilots have delivered good value for money, with net annual benefits expected to increase over time through declining support intensity and growing health and criminal justice savings.

KEY DATA SUMMARY

Indicator	Value
Total people receiving support	1,387
Total people housed	1,061
Total expenditure (Dec 2022)	£27.6 million + £0.4m in-kind
Pilot budgets	GM £8.0m; Liverpool £7.7m; WM £9.6m
Extension funding (2022)	£13.9 million
Annual unit cost per person supported	£7,737 (range £5,558–£12,613)
Cumulative cost per person housed	£26,348 (range £18,430–£43,841)
Expected annual benefit per person	£15,880

Observed annual benefit at 12 months	£8,429
BCR (expected)	2.1:1
BCR (12-month observed)	1.1:1
Deadweight assumption	30%
Negative outcome adjustment	15%
Graduations by Dec 2022	86
Deaths on programme	90

SOURCES

- ICF Consulting (2024). *Evaluation of the Housing First Pilots: Cost Benefit Analysis — Final Report*. Ministry of Housing, Communities and Local Government. October 2024. Available at: https://assets.publishing.service.gov.uk/media/671a6fea603993b7a8f75db5/Housing_First_Cost_Benefit_Analysis_Report.pdf
- MHCLG (2024). *Evaluation of the Housing First Pilots: Final Synthesis Report*. October 2024. Available at: https://assets.publishing.service.gov.uk/media/671a70221898d9be93f75db4/Housing_First_Final_Synthesis_Report.pdf
- Pleace, N. and Culhane, D. (2016). *Better than Cure? Testing the case for enhancing prevention of single homelessness in England*. London: Crisis.
- Centre for Social Justice (2021). *Housing First: Housing Led Solutions to Rough Sleeping*. London: CSJ.

CASE STUDY 4: Social Return on Investment (SROI)

Table 11 - SROI Analysis Programme Snapshot

Element	Description
Programme	Employment and Training Support for Unemployed Westminster Residents
Location	Westminster, London, UK
Scale	Medium — £96,931 total contract value, 188 participants, 36 into employment
Duration	April 2009 – September 2011 (30 months)
Target Population	Unemployed adults (working <16 hours/week, not in full-time education) in Westminster
Funding	Tailored, participant-centred support pathway delivered by housing association, contrasting with generic one-size-fits-all employment programmes
Key Innovation	£96,931 total contract; funders: London Councils, European Social Fund, Westminster Council

THE CHALLENGE

Westminster, despite being one of London's wealthiest boroughs, had persistent pockets of unemployment and deprivation. The Octavia Foundation programme was part of a wider Westminster Works project involving a consortium of housing associations, charities, and social enterprises tackling borough-wide unemployment. The wider project helped 1,300 people access training and paid work.

Mainstream employment programmes offered standardised approaches that participants described as rigid, requiring attendance at sessions covering skills they already possessed. The Octavia Foundation sought to design a programme tailored to individuals rather than driven by output targets.

THE INTERVENTION

The programme was delivered by one full-time Employment Advisor, project-managed by the Octavia Foundation's Community Initiatives Manager. The contract specified a minimum of six hours' contact per client, with output targets for enrolments, training, volunteering, work placements, and jobs secured.

Core features included: an initial group induction followed by individual induction sessions to agree action plans; CV development tailored to specific work types; interview workshops; weekly job clubs to research opportunities; and follow-up support after



employment for up to nine months. Volunteering, training, and work placement options were available depending on individual needs.

The programme prioritised its tailored approach over rigid adherence to contact-hour targets — in practice, some clients received more or less than six hours depending on need. Most referrals came from community organisations including the Harrow Road Partnership, Kensington Volunteer Centre, and housing associations.

WHY SOCIAL RETURN ON INVESTMENT WAS CHOSEN

The Octavia Foundation wanted to demonstrate the value of its tailored approach compared to mainstream employment programmes. SROI was selected as a recognised method for measuring social value relative to resources invested, following the principles established by the SROI Network and the New Economics Foundation (NEF). The methodology enabled the Foundation to capture the direct financial benefits of moving people into employment alongside estimated health benefits, whilst maintaining a conservative approach that excluded harder-to-quantify outcomes such as family wellbeing and community impacts.

HOW THE SOCIAL RETURN ON INVESTMENT WAS CONDUCTED

Evaluator: Kam Chung, Head of Service Development at Octavia Housing (internal evaluation)

Stakeholder engagement: The evaluator reviewed funder impact reports, interviewed the Employment Advisor delivering the project, and interviewed five people (out of 36 helped into work) who participated in the programme. The evaluation examined participants' circumstances before and after the programme, including work status, income sources, sense of wellbeing, and aspirations.

SROI calculation approach:

The evaluation included only direct benefits of moving people into employment and estimated health benefits, deliberately excluding indirect benefits (family wellbeing, increased purchasing power, community effects) to avoid over-claiming. Specifically:

- **Benefits saved:** Calculated using the mean average of the lowest Jobseeker's Allowance rate for a single person and the JSA rate for a couple, plus Housing Benefit based on average London social housing rents (National Housing Federation data).
- **Part-time work adjustment:** 17% deducted for estimated top-up benefits claimed whilst in part-time work (borrowed from Ready for Work SROI in the absence of programme-specific data).
- **Tax contributions:** 16% added for estimated Income Tax and National Insurance contributions (also borrowed from Ready for Work SROI).



- **Health benefits:** £508 per person into work per year, using the DWP Cost-Benefit Analysis Framework proxy (no uplifts applied).
- **Deadweight:** 15% discounted for what might have occurred without the intervention.
- **Attribution:** 20% discounted for the possible contribution of other agencies.
- **Drop-off:** 50% per year over a five-year projection (standard assumption used across comparable SROIs).
- **Financial proxy sources:** All proxies taken from government sources where available (DWP benefit rates, DWP CBA Framework). Deadweight, attribution, and drop-off assumptions borrowed from the Ready for Work SROI (Business in the Community, 2012).

KEY FINDINGS

Programme outputs:

- 188 people enrolled; 36 helped into work
- 28 sustained employment at six months (against a target of 19)
- Programme met or exceeded all output targets for training, volunteering, and placements

SROI ratio: £4.12 of social value for every £1 invested.

The total calculated social impact over five years was £399,357 against an investment of £96,931. This included welfare benefit savings, tax contributions, and estimated health cost savings to the NHS. The ratio reflects only direct financial benefits and health benefits; the actual social value created was likely higher given excluded indirect benefits.

Participant experiences (pseudonyms, from SROI report):

The evaluation documented four participant journeys illustrating the programme's impact:

- *Laura* (28, single): Former full-time carer for her father, five years out of work. After a year in the programme including action planning, CV workshops, mock interviews, and confidence coaching, she secured work as an Arabic translator at Heathrow Airport.
- *Tomaz* (married, one daughter): Nine years out of work with a history of substance misuse. After six one-to-one sessions focusing on CV and interview preparation, he secured an apprenticeship at St Mungo's Housing, progressing to a permanent post. His wife also found work.
- *Tamara* (41, single mother): Made redundant after ten years with the same employer. After previous rigid employment programmes proved unhelpful, she

contacted 30 housing associations seeking volunteer placements, receiving one response — from the Octavia Foundation. After three months and several one-to-one sessions, she secured a volunteer placement, progressing to a paid full-time post in resident liaison.

- *Paulo* (34, married): Periodic retail work since finishing studies. After attending workshops and one-to-one sessions over three months, he found retail work and progressed through three jobs, reaching the Ferrari sales team within 18 months.
- All interviewees spoke about positive impacts to confidence, self-esteem, health, and family life beyond what could be captured in the SROI ratio. The tailored approach was mentioned positively by all interviewees, with some contrasting it with the rigidity of other employment programmes they had experienced.

Subsequent impact:

The Octavia Foundation continued delivering employment and training through its successor programme (Future Foundations), funded by Octavia Housing, with specific targets for helping Octavia Housing residents into employment.

KEY DATA SUMMARY

Element	Detail
Evaluator	Internal — Kam Chung, Head of Service Development, Octavia Housing
SROI type	Evaluative (retrospective)
Contract value	£96,931
Participants	188 enrolled; 36 into employment; 28 sustained at 6 months
Stakeholders consulted	Employment Advisor; 5 programme participants (of 36 employed)
SROI ratio	£4.12 : £1
Total social impact (5 years)	£399,357
Deadweight	15%
Attribution	20%
Drop-off	50% per year
Health proxy	£508/person/year (DWP CBA Framework)
Perspective	Conservative — direct benefits and health only; excluded family, employer, community benefits



SOURCES

- Octavia Foundation (2012). *Placing a Value on Work: A Social Return on Investment Report*. London: Octavia Foundation. Available at: https://www.octaviafoundation.org.uk/assets/0000/1500/SROI_Report_Guardian_Version.pdf
- Nicholls, J., Lawlor, E., Neitzert, E., & Goodspeed, T. (2012). *A Guide to Social Return on Investment*. SROI Network.
- Business in the Community (2012). *Social Return on Investment of Ready for Work*.
- Department for Work and Pensions (2008). *Cost Benefit Analysis Framework for Employment Programmes*.

CASE STUDY 5: Mixed Foundation Methods – Process and Outcome Evaluation

Table 12 - Mixed Methods Analysis Programme Snapshot

Element	Description
Programme	Orientation programme for refugee health professionals' labour market integration
Location	Hamburg, Germany (University Medical Centre Hamburg-Eppendorf)
Scale	Small — 29 participants across 2 cohorts
Duration	3-month programme per cohort (2018–2019 cohorts evaluated), with 3-month follow-up
Target Population	Refugee health professionals (all health professions) registered in Hamburg, applying for or preparing to apply for qualification recognition in Germany
Key Innovation	Integrated orientation combining German medical terminology, cross-cultural coaching, and hands-off job shadowing (3 days/week) without requiring work permits
Budget	European Social Fund funded (4-year pilot programme). Participants neither paid for participation nor paid to participate, except for travel costs and food expenses

THE CHALLENGE

Refugee health professionals in Germany face multiple structural and psychological barriers preventing re-entry to their professions despite acute healthcare workforce shortages. Access to the labour market and language courses often remains restricted. Health professions are highly regulated, and the recognition of prior qualifications, lack of proof of credentials, and unfamiliarity with the German healthcare system pose significant barriers. The experience of discrimination, loss of professional identity, and loss of self-confidence create additional psychological obstacles. The longer refugees remain out of their professional field, the harder it becomes to re-enter — leading to de-professionalisation and skill erosion.

Existing programmes for international health professionals largely lacked methodological quality and transparency in evaluation, making it unclear how effective they were. Most programmes addressed international medical graduates specifically, with few targeting all health professions.

THE INTERVENTION

The programme lasted three months and comprised three modules:

German medical terminology course (2 days/week): Focused on medical communication situations, case studies, and specialised vocabulary. Supported through blended learning with an online platform providing interactive video and audio exercises. Emphasised active use rather than passive comprehension.

Cross-cultural coaching (1 day/week): Addressed cross-cultural aspects of clinical work including communication, interprofessional work, confidentiality, and patient relationships. Included strategies for dealing with difficult situations such as racism, conflict resolution, and job interview preparation through role plays.

Job shadowing at University Medical Centre (3 days/week): Hands-off observation only (no work permits required). Participants matched towards their qualifications. Observation of ward rounds, clinics, and procedures; interaction with healthcare staff; exposure to German clinical practice.

Participants were matched to shadowing placements corresponding to their qualifications, though in practice this proved challenging — particularly for non-physician professions such as dentists, dermatologists, and psychologists.

WHY FOUNDATION METHODS WERE CHOSEN

As a small-scale pilot (29 participants across two cohorts), the programme could not justify expensive comparison-group designs or cost-benefit analysis. No validated instruments existed for this specific context. A mixed-methods design combining outcome monitoring (quantitative questionnaires) and stakeholder feedback (qualitative interviews) was chosen to capture both measurable skills improvement and the nuanced experiences of participants navigating complex integration pathways.

The evaluation followed the Kirkpatrick Training Evaluation framework, addressing: (1) participant satisfaction (reaction); (2) knowledge and skills gained (learning); and (3) broader impacts on personal situation and career prospects.

HOW THE FOUNDATION METHODS EVALUATION WAS CONDUCTED

Evaluators: Sidra Khan-Gökkaya (PhD student, Department of Medical Psychology, University Medical Centre Hamburg-Eppendorf) and two research assistants (psychology students), all with experience in migration research.

Quantitative instruments (outcome monitoring):

Weekly process questionnaires (self-developed, plain language German): Assessed satisfaction and topic relevance for each programme component using 4-point Likert

scales with open-response questions. 251 feedback forms collected for the German course (out of 360 possible); 217 for coaching (out of 348 possible).

Post-programme outcome questionnaire (self-developed, based on systematic review of qualification programmes for immigrant health professionals): 16 items on 5-point Likert scale assessing improvement across three categories — language skills, professional skills, and formal resources/career prospects. Translated into Arabic and Farsi to minimise language barriers. Completed by 26 of 29 participants (3 non-respondents: 2 left early, 1 severe language barriers).

Qualitative instruments (stakeholder feedback):

Participant interviews: 24 participants interviewed at four time points (pre-programme, half-time, post-programme, 3-month follow-up). Interviews lasted 4–38 minutes (median 18 minutes). Conducted in German.

Staff interviews: 3 staff (2 teachers, 1 practical instructor) interviewed at three time points.

Total: 105 interviews across participants and staff.

Analysis: Interviews transcribed verbatim by professional agency. Qualitative content analysis (Mayring framework) using MAXQDA version 10. Deductive categories derived from interview guides; inductive categories emerged from data. Two research assistants independently coded 25% of all interviews; uncertainties discussed. Results presented in interdisciplinary research colloquium with migration research experts.

KEY FINDINGS

Satisfaction with programme components:

German medical terminology course: Mean satisfaction 3.37/4. Average attendance 70% (21 of 30 participants per session across both cohorts).

Cross-cultural coaching: Mean satisfaction 3.58/4 (rated 'very satisfied'). Topics rated on average as 'very important' (mean 3.55/4). Average attendance 62%.

Post-programme outcomes (5-point Likert, from published Table 3):

Outcome	Mean	SD
Learned new technical language	4.08	0.76
Less afraid of speaking German while working	4.24	0.78
Encouraged to work in job again	4.22	1.09
Know more about the healthcare system	4.17	1.17
Know more about working culture in German hospitals	4.09	1.08
Feel better prepared to work in Germany	4.04	0.71
Got good insight into German hospitals	4.04	0.98
More self-confident while working	3.96	1.15
German improved through the project	3.79	0.83
Gained new medical knowledge	3.96	0.83
Built professional networks	2.96	1.22
Found a job	1.35	0.81

Qualitative findings — programme strengths:

- Participants valued the division between language course, coaching, and shadowing, describing how observations during shadowing could be discussed in subsequent classes. The German terminology course was praised for its practice-oriented teaching, including simulated patient consultations. Cross-cultural coaching addressed not only professional contexts but also everyday life in Germany — participants found strategies for dealing with racism particularly valuable.
- Language confidence emerged as the strongest outcome: the highest-rated item ('less afraid of speaking German while working', 4.24/5) was strongly supported by qualitative accounts. As one participant described: *'Before, I was too scared to speak German with patients. Now I know I can try'* (P62, translated from German).
- Professional identity restoration was a powerful qualitative finding not fully captured by quantitative measures. Participants reconnected with their professional selves after years of de-professionalisation. Positive feedback from ward colleagues during shadowing was described as encouraging.

Qualitative findings — programme weaknesses:

- Job shadowing supervision was inconsistent and emerged as the most significant implementation problem. Participants assigned to well-supervised placements reported positive outcomes; those without designated supervisors reported boredom, disappointment, and in some cases reduced self-confidence. Some participants were not assigned placements at all, assigned very late (six weeks after programme start), or assigned to wards not matching their qualifications. As one participant stated: *'No one helps me finding orientation. I come every morning at six o'clock and I try to do something — but no one helps me. I feel like a stranger at the ward'* (P26).
- Group composition was challenging: participants had heterogeneous language competencies and professional backgrounds. Non-physicians (psychologists, dentists, physiotherapists) felt content focused disproportionately on physician-relevant topics. Financial barriers (travel costs of approximately €160/month) prevented full participation for some, despite the programme itself being free.
- Staff experienced difficulty finding shadowing placements across diverse specialties and observed participant demotivation when shadowing was poorly supervised. They proposed establishing mentoring programmes and matching participants with medical students as tandem partners.

Follow-up outcomes (3 months post-programme, qualitative):

Most outcomes reported at programme end were sustained at three-month follow-up. However, negative effects from poorly supervised shadowing were also sustained. Some participants maintained contact with shadowing hospitals and established learning groups. Employment outcomes were limited — consistent with the 2–5 year licensing timeline for health professions in Germany.

KEY DATA SUMMARY

Element	Detail
Evaluators	Internal — PhD student + 2 research assistants, Dept. of Medical Psychology, UKE Hamburg
Sample	29 participants (Cohort 1: 17; Cohort 2: 12) + 3 staff
Participant origins	Syria (21), Afghanistan (2), Algeria (2), Iraq, Libya, Congo, Iran, Moldova
Participant sex	18 male (62%), 11 female (38%)
Participant professions	Physician (21), Psychologist (2), Medical technical assistant (2), Physiotherapist, Dentist, Pharmacist, Nurse
Data collected	251 weekly questionnaires (German course); 217 (coaching); 26 post-programme questionnaires; 105 interviews

Interview duration	4–38 minutes (median 18 minutes)
Transcription	Professional agency (verbatim)
Analysis	Qualitative content analysis (Mayring) using MAXQDA 10
Quality assurance	25% independently coded by 2 research assistants; interdisciplinary validation in research colloquium
Limitations acknowledged	Self-developed instruments (not validated); self-assessment bias; small sample (not generalisable); social desirability bias; language barriers excluding some participants
Evaluation cost	Conducted within existing university research capacity (no separate budget)

SOURCES

- Khan-Gökkaya, S. and Mösko, M. (2020). 'Process- and outcome evaluation of an orientation programme for refugee health professionals.' *Medical Education Online*, 25(1), 1811543. <https://doi.org/10.1080/10872981.2020.1811543>
- Khan-Gökkaya, S., Higgen, S., & Mösko, M. (2019). Qualification programmes for immigrant health professionals: A systematic review. *PLoS ONE*, 14(11), e0224933. <https://doi.org/10.1371/journal.pone.0224933>
- Kirkpatrick, D.L. (1994). *Evaluating Training Programs: The Four Levels*. Berrett-Koehler.
- Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Beltz Verlag.

Annexes



ANNEXES

ANNEX STRUCTURE

Glossary of Key Evaluation Terms

Annex A: Data Management, GDPR, Cross-Border Data Protection (A.1-A.4)

Annex B.1: Theory of Change Templates

Annex B.2: Outcome Monitoring Templates

Annex B.3: Stakeholder Feedback Templates

Annex B.4: Cost-Effectiveness Analysis Templates

Annex B.5: Multi-Criteria Decision Analysis Templates

Annex B.6: SROI

Annex B.7: CBA

Annex B.8: Advanced Causal & Complexity Evaluation Templates [QED, RCT, Realist Evaluation in order]

Annex B.9: EU Integration Templates

GLOSSARY OF KEY EVALUATION TERMS

This glossary defines technical terms used throughout the Green Book. Terms are explained in plain language for practitioners who may not have formal research or evaluation training.

Table 13 - Glossary of Key Evaluation Terms

Term	Definition & Example
Attrition	When participants drop out of a programme or stop providing data before the evaluation is complete. High attrition can bias results if those who drop out differ systematically from those who stay. <i>Example:</i> If 100 people start a job training programme but only 60 complete the final survey, 40% attrition has occurred. If those who dropped out were the ones struggling most, the results will overstate programme success.
Baseline	A measurement taken before the programme starts, used as a reference point to assess change. <i>Example:</i> Measuring participants' wellbeing scores at the start of a counselling programme, so that scores at the end can be compared to see if improvement occurred.
Causal inference	Methods for determining whether a programme actually caused observed changes, rather than changes happening for other reasons.

	Example: <i>Did participants get jobs because of the training programme, or would they have found jobs anyway as the economy improved?</i>
Comparison group	A group of people similar to programme participants who did not receive the programme, used to estimate what would have happened without the intervention. Example: <i>Comparing employment rates of training programme participants with similar job seekers who were on a waiting list.</i>
Confounding variable	A factor other than the programme that could explain observed outcomes, potentially leading to incorrect conclusions about programme effectiveness. Example: <i>If a youth mentoring programme operates only in affluent areas, neighbourhood wealth (not the mentoring) might explain better outcomes.</i>
Cost-effectiveness	The cost required to achieve one unit of a specific outcome. Enables comparison of which approach achieves a goal most efficiently. Example: <i>Programme A costs €2,000 per participant gaining employment; Programme B costs €3,500. Programme A is more cost-effective for this outcome.</i>
Counterfactual	What would have happened to participants if the programme had not existed. The fundamental question in impact evaluation. Example: <i>Without the housing support programme, how many of these families would still have become homeless?</i>
Deadweight	The proportion of outcomes that would have occurred even without the programme. Must be estimated and subtracted from programme effects. Example: <i>If 30% of training programme participants would have found jobs without the programme, 30% deadweight is applied.</i>
Discounting	Adjusting the value of future costs and benefits to reflect the fact that people generally value things received now more than identical things received in the future. Example: <i>€1,000 of benefits occurring in 10 years is worth less in today's terms than €1,000 of benefits occurring now. At a 3% discount rate, €1,000 in 10 years is worth approximately €744 today.</i>
Drop-off	The rate at which programme outcomes diminish over time after the programme ends. Example: <i>Employment programme benefits may be strong in Year 1 (90% still employed) but decline by Year 3 (60% still employed).</i>
Financial proxy	A monetary value assigned to an outcome that does not have a market price, used in SROI to estimate social value in monetary terms.

	<p>Example: The value of “reduced social isolation” might be estimated using the cost of social activities that would achieve a similar level of social connection.</p>
Indicator	<p>A specific, measurable variable used to track whether an outcome is occurring.</p> <p>Example: For the outcome “improved mental health,” an indicator might be scores on the Warwick-Edinburgh Mental Wellbeing Scale.</p>
Logic model	<p>A simplified visual representation showing the linear sequence from programme inputs to activities, outputs, outcomes, and impacts. Less detailed than a Theory of Change.</p> <p>Example: A one-page diagram showing: Funding + Staff → Training sessions → Certificates awarded → Jobs gained → Economic independence.</p>
Monetisation	<p>Converting non-financial outcomes into monetary values to enable comparison with costs.</p> <p>Example: Estimating that each case of homelessness prevented saves €18,000 per year in public service costs.</p>
Outcome	<p>A change experienced by programme participants or other stakeholders as a result of the programme. Distinct from outputs (which count what was delivered).</p> <p>Example: Employment gained, improved health, increased confidence (outcomes) vs. 50 training sessions delivered, 200 people attended (outputs).</p>
Output	<p>A direct, countable product of programme activities. Outputs describe what the programme did, not what changed for participants.</p> <p>Example: 120 participants completed training (output). Whether those participants gained employment is an outcome, not an output.</p>
Proportionality	<p>The principle that evaluation effort and sophistication should match the scale of the programme and the importance of the decisions the evaluation will inform.</p> <p>Example: A €50,000 community garden project does not need a Randomised Controlled Trial; a €50 million national welfare reform does.</p>
Selection bias	<p>Systematic difference between people who participate in a programme and those who do not, which can distort conclusions about programme effectiveness.</p> <p>Example: Participants in a voluntary job training programme may be more motivated than non-participants, making the programme appear more effective than it is.</p>

Sensitivity analysis	<p>Testing how conclusions change when key assumptions are varied. Demonstrates robustness of findings.</p> <p>Example: <i>If the CBA shows positive results assuming 5% deadweight, does it remain positive at 20% deadweight? At 40%?</i></p>
Stakeholder	<p>Any individual, group, or organisation affected by or with an interest in the programme and its evaluation.</p> <p>Example: <i>Participants, families, programme staff, funders, partner organisations, local government, community members.</i></p>
Theory of Change	<p>A detailed explanation of how and why a programme is expected to produce its intended outcomes, including causal pathways, mechanisms, assumptions, and contextual factors. More comprehensive than a logic model.</p> <p>Example: <i>Not just “training → employment” but specifying that training builds specific skills and confidence and professional networks, provided labour market demand exists and participants can overcome transport and childcare barriers.</i></p>
Validated measure	<p>A measurement tool (questionnaire, scale, assessment) that has been scientifically tested and shown to reliably and accurately measure what it claims to measure.</p> <p>Example: <i>The PHQ-9 is a validated measure of depression severity — it has been tested across populations and shown to consistently measure depressive symptoms. A homemade survey asking “are you happy?” is not validated.</i></p>

Terms are listed alphabetically. For detailed methodological guidance on any of these concepts, refer to the relevant chapter and section indicated in the main text.

ANNEX A

A.1: Simple Data Management Plan (DMP) Template

Purpose: This is a non-technical, internal planning tool. Use it *before* your project starts to get your team on the same page. It ensures you only collect what you need and handle it safely.

Template 1 - Simple Data Management Plan (DMP) Template

Section	Guiding Questions (Fill in your answers)
1. Project & Purpose	<p>What is this evaluation for? (e.g., <i>Annual report for our funder? Internal service improvement?</i>)</p> <p>What are the 1-3 key questions we need to answer? (e.g., <i>Did our clients' well-being improve? Are they satisfied?</i>)</p>
2. Data to be Collected	<p>What data will we collect? (Check all that apply)</p> <p><input type="checkbox"/> Participant Demographics</p> <p><input type="checkbox"/> Service Participation (e.g., attendance)</p> <p><input type="checkbox"/> Outcome Data (e.g., pre/post surveys)</p> <p><input type="checkbox"/> Stakeholder Feedback (e.g., satisfaction)</p>
3. Key Indicators (from ToC)	<p>What are the 3-5 most important indicators from our Theory of Change (ToC) that we will track?</p> <p>1. (e.g., <i>Change in well-being score</i>)</p> <p>2. (e.g., <i>% employed at 6 months</i>)</p> <p>3. (e.g., <i>Average satisfaction score</i>)</p>
4. Collection & Storage	<p>How will we collect it? (e.g., <i>Admin data from intake forms, paper surveys at exit, phone interviews at 6-month follow-up</i>).</p> <p>Where will it be stored? (e.g., <i>A single password-protected Excel file on a secure shared drive</i>).</p>

5. Roles & Responsibilities	<p>Who is responsible for:</p> <ul style="list-style-type: none"> * Collecting it? (e.g., <i>Case managers during intake</i>) * Entering it? (e.g., <i>Admin assistant, weekly</i>) * Checking it for quality? (e.g., <i>Programme manager, monthly</i>) * Analysing & reporting it? (e.g., <i>Programme manager, quarterly</i>)
6. Ethics & GDPR	<p>What is our legal basis? (e.g., <i>Informed Consent</i>).</p> <p>How will we ensure confidentiality? (e.g., <i>All data will be anonymised in the Excel file using a Participant ID. The "key" linking IDs to names will be kept in a separate, encrypted file.</i>)</p> <p>When will it be deleted? (e.g., <i>Anonymised data kept for 5 years; personal/identifiable data deleted 1 year after project end</i>).</p>

A.2: GDPR-Compliant Consent Form Template

Purpose: This is a *template*. You must adapt it to your specific project and have it reviewed by your organisation's Data Protection Officer or legal advisor. Use clear, simple language.

Provider's Note: Under GDPR, you must have a "legal basis" for processing data. If you are *not* using **Consent** (e.g., you are using **Legitimate Interest** or **Public Task**), you do not need this form, but you *must* still provide participants with a clear **Privacy Notice** explaining what you are doing.

[Your Organisation Logo/Name]

Participant Consent Form for Evaluation

We are asking for your permission to use your information for programme evaluation. This helps us understand what we are doing well and how we can improve our service for everyone.

- **Who we are:** [Your Organisation Name]
- **What this is for:** We are collecting information to evaluate the [Name of Programme]. Your answers will help us learn and report to our funders.

What information will you collect? We will collect:

- Your answers to our surveys (about your well-being, skills, etc.).
- Information about your use of our service (like your start date and the number of sessions you attended).
- Basic demographic information (like your age and gender).

How will you use my information and keep it safe?

- Your data will be combined with answers from other participants to create statistics and reports.
- **You will not be personally identified.** Your name will be removed and replaced with a code number (anonymisation).
- All data is stored securely on password-protected computers.
- Only the direct evaluation team will have access to the data.

Your Rights Your participation is **100% voluntary**.

- You can **refuse** to answer any question.
- You can **stop** at any time.
- You can **withdraw** your consent at any time by contacting [Email/Phone], and we will delete your data.
- Refusing to take part will **not** affect the service you receive from us in any way.
- You have the right to access or correct your data.

Statement of Consent (Please tick the boxes and sign)

I have read and understood this form. I have had the chance to ask questions. I agree to my data being used for evaluation as described above.

Name: _____

Signature: _____

Date: _____



A.3: Data Quality Checklist

Purpose: Use this simple checklist for a quick, regular (e.g., quarterly) review of your data system to catch problems early.

Template 2 - Data Quality Checklist

Quality Dimension	Action / Check	Status (OK / Needs Fix)
1. Completeness	Are baseline forms being 100% completed <i>at intake</i> (before the service starts)?	
	Are exit forms being collected from most participants? (Target >70%)	
	If we do follow-up , is our response rate acceptable? (Target >60%)	
2. Accuracy	Have we spot-checked 5 random paper forms against the spreadsheet to check for data entry errors?	
	Are there any obvious outliers or strange numbers (e.g., an age of "120" or a score of "11" on a 1-10 scale)?	
3. Consistency	Are all staff using the <i>exact same</i> forms and definitions for data collection?	
	Are data codes consistent? (e.g., is "Female" always coded as "F" and not sometimes "1" or "Fem"?)	
4. Timeliness	Is data being entered promptly (e.g., within 1 week of collection) or is there a backlog?	
5. Process	Have all <i>new</i> staff involved in data collection been trained on these procedures?	
	Are all data files stored securely (password-protected, locked cabinet)?	
	Are we discussing these data quality checks in our regular team meetings?	

A.4: Cross-Border Data Transfer Checklist

Template 3 - Cross-Border Data Transfer Checklist

Section	Guiding Questions / Required Documentation	Status (Completed /In Progress)	Notes
1. Assessment	Is a Third Country Partner (outside EEA) receiving data?		
	Does that country have an Adequacy Decision ? (If yes, no SCCs usually needed)		
	Is the data highly sensitive (e.g., health, criminal justice)? (Requires stricter safeguards)		
2. Legal Basis	Has the Data Management Plan (DMP) identified the specific data being transferred?		
	Has the legal basis for processing (e.g., explicit consent, public task) been confirmed for this specific transfer?		
3. Documentation	Have Standard Contractual Clauses (SCCs) been fully signed by both the EU partner (exporter) and the Third Country Partner (importer)?		
	Has a Transfer Impact Assessment (TIA) (or equivalent) been completed to ensure local laws do not conflict with GDPR standards?		
	Have participants given explicit consent for international data transfer, naming the receiving country/partner?		
4. Technical Safeguards	Is the data pseudonymised/anonymised before transfer where possible?		
	Are technical security measures (e.g., encryption, password protection) applied during transfer?		

ANNEX B

B.1: THEORY OF CHANGE TEMPLATES AND TOOLS

B.1.1 ToC Narrative Template

PROGRAMME OVERVIEW (1 page)

- Target population and eligibility criteria
- Service description and core activities
- Geographic coverage and delivery settings
- Annual reach and service intensity
- Intended ultimate outcomes

LONG-TERM GOAL & PRECONDITIONS (1-2 pages) Ultimate impact statement:
[Describe the fundamental change you aim to contribute to, acknowledging multi-causal nature]

Key preconditions required:

1. [Precondition 1]
 - Why this is necessary
 - How programme addresses it
2. [Precondition 2]
3. [Precondition 3] [etc.]

Evidence base:

- Research evidence supporting these preconditions
- Practice wisdom from experienced practitioners
- Lived experience insights from service users

CAUSAL PATHWAYS (2-3 pages)

Pathway 1: [Name, e.g., "Skills Development Pathway"]

- Programme activities involved
- Immediate outputs expected
- Short-term outcomes (0-6 months)
- Medium-term outcomes (6-24 months)
- Contribution to long-term impact
- Key assumptions underpinning this pathway

Pathway 2: [Name] [Same structure]

Pathway 3: [Name] [Same structure]

Interactions between pathways: [Describe how pathways reinforce each other, potential synergies, any tensions]

KEY ASSUMPTIONS (1-2 pages)

Critical assumptions by causal stage:

Activities → Outputs:

- Assumption 1: [Statement]
 - Evidence supporting: [Research, practice wisdom, or logic]
 - Evidence questioning: [Contrary evidence or risks]
 - Plan to test: [How you'll verify this assumption]

Outputs → Short-term outcomes:

- Assumption 2: [Statement]
 - [Same structure]

Short-term → Medium-term outcomes:

- Assumption 3: [Statement]

Medium-term outcomes → Long-term impacts:

- Assumption 4: [Statement]

Cross-cutting assumptions:

- Assumption about target population engagement
- Assumption about external environment stability
- Assumption about implementation fidelity

CONTEXTUAL FACTORS (1 page)

Economic factors:

- [e.g., Local labour market conditions, benefit system rules]
- How they influence programme effectiveness
- Monitoring and adaptation strategy

Policy and regulatory factors:

- [e.g., Relevant policy frameworks, regulatory requirements]
- Influence on programme operations
- Adaptation strategy

Social and cultural factors:

- [e.g., Stigma, community norms, cultural appropriateness]
- Impact on participant engagement and outcomes
- Mitigation strategies

Geographic and infrastructural factors:

- [e.g., Urban/rural setting, transport, service availability]
- Effects on access and complementarity
- Adaptation approaches

VISUAL THEORY OF CHANGE (1 page) [Insert diagram here]

Design guidance:

- Use landscape A3 or A4 orientation
- Colour code different pathways
- Boxes for programme elements, ovals/clouds for assumptions
- Clear arrows showing causal direction
- Context factors shown as surrounding/influencing system
- Legend explaining visual conventions
- Readable font sizes (minimum 10pt)

IMPLICATIONS FOR EVALUATION (1-2 pages)

Key evaluation questions derived from ToC:

1. Process questions:
 - Are we implementing activities as intended?
 - Are we reaching target population?
 - What implementation challenges arise?
2. Outcome questions:
 - Are participants experiencing expected short-term outcomes?
 - Do medium-term outcomes materialise?
 - Which causal pathways appear strongest?
3. Assumption testing questions:
 - Is Assumption X holding true?
 - Where are our assumptions being challenged?

Measurement implications:

ToC Element	Indicator(s)	Data Source	Timing	Responsible
Output 1	[Specific indicator]	[Source]	[When]	[Who]
Outcome 1	[Indicator]	[Source]	[When]	[Who]
Assumption 1	[How to test]	[Source]	[When]	[Who]

Data collection approach:

- Administrative data systems
- Participant surveys/assessments
- Staff reporting
- Partner feedback
- Case studies/qualitative research

Analysis plan:

- Frequency of analysis (monthly, quarterly, annually)
- Disaggregation by key characteristics
- Trend analysis over time
- Comparison across sites/cohorts if applicable

Use of findings:

- Service improvement processes
- Reporting to stakeholders
- ToC refinement and updating

B.1.2 Assumption Testing Matrix

Template 4 – Theory of Change Assumption Testing Matrix

Assumption	Evidence For	Evidence Against	How to Test	If False, Then...	Priority (H/M/L)
Example: "Participants who complete 80%+ of training sessions will demonstrate measurable skills improvement"	- Research on adult learning shows dosage effects- Pilot data showed correlation	- High attrition in first pilot- Skills transfer depends on quality not just quantity	- Track attendance vs. skills assessment scores- Pre-post skills tests- Staff observations	- Revise attendance expectations- Focus on engagement quality- Add motivational support	H
[Assumption 1]					
[Assumption 2]					

Priority scoring:

- High (H): Critical to programme success AND highly uncertain
- Medium (M): Important but moderately certain, or less critical but very uncertain
- Low (L): Either not critical or already well-evidenced

B.1.3 Stakeholder Consultation Protocol

Purpose: Develop Theory of Change through participatory process involving diverse stakeholders

Participants to invite:

- Front-line staff
- Management/leadership
- Service users or representatives (ideally 20-30% of group)
- Key partners/referrers
- Funder representatives (if appropriate at this stage)
- Board members or governance representatives

Workshop structure:

Part 1: Introduction

- Explain ToC purpose and approach
- Overview of workshop process
- Ground rules for inclusive participation

Part 2: Long-term goal identification

1. Individual reflection: What ultimate change are we trying to achieve?
2. Small group sharing and consolidation
3. Plenary discussion reaching consensus

Part 3: Preconditions mapping

1. For agreed long-term goal(s), ask: "What needs to happen first?"
2. Capture preconditions on sticky notes
3. Arrange on wall in rough causal sequence
4. For each precondition, repeat: "What needs to happen for this?"
5. Continue until reaching programme activities
6. Identify different pathways if multiple routes to change

Part 4: Assumptions identification

1. For each key causal link, ask: "What must be true for X to lead to Y?"
2. Capture assumptions on different coloured sticky notes
3. Map assumptions onto causal diagram
4. Prioritise: Which assumptions are most critical? Most uncertain?

Part 5: Context factors

1. Brainstorm external factors affecting programme
2. Categorise: Economic, policy, social, geographic, temporal
3. Discuss: How do these enable or constrain our work?

Part 6: Next steps

1. Agree who will create visual ToC and written narrative
2. Set timeline for draft review
3. Plan for validation and refinement

B.1.4 Visual ToC Design Guidance

Software options:

Simple presentations:

- PowerPoint / Google Slides: Adequate for most programmes, familiar to staff

Dedicated ToC tools:

- Theory of Change Tech (toctech.org): Free, designed specifically for ToC

Visual mapping tools:

- Kumu (kumu.io): Sophisticated network mapping
- Lucidchart, Draw.io: Flexible diagramming tools
- Miro: Collaborative whiteboard

B.1.5 Theory of Change Quality Assurance Protocol

PART A: INITIAL TOC QUALITY ASSESSMENT

Use when ToC first developed or received from partners.

Causal Logic

- Each link (inputs → activities → outputs → outcomes → impacts) clearly articulated
- Activities distinguished from outcomes (what staff DO vs. what CHANGES achieved for participants)
- Outcome timeframes realistic (short: 0-6 months, medium: 6-24 months, long: 24+ months)

Testability of Assumptions

- Assumptions specific not vague (e.g. NOT "training works" BUT "participants completing 80%+ sessions will demonstrate measurable skills improvement")
- Assumptions prioritised: High (critical + uncertain), Medium (important or uncertain), Low (not critical or well-evidenced)
- Testing methods specified for each high-priority assumption

Appropriate Complexity

- Complexity matches programme sophistication
- ToC usable: Visual fits one page (A3 max), narrative 3-8 pages, staff can explain without documents
- Context factors acknowledged (economic, policy, social, geographic)

Visual Quality

- Clear, readable
- Consistent shapes for element types
- Accessible to non-experts

Stakeholder Buy-In

- Developed participatively
- Reflects actual current programme (not aspirational or historical)
- Owned by team (referenced in decisions, displayed, used in induction)

PART B: ANNUAL REVIEW PROCESS

Conduct annually PLUS after major shocks (policy change, pandemic, unexpected findings).

Step 1: Evidence Review

- New research on similar programmes or causal mechanisms
- Own evaluation findings (outcome monitoring, impact evaluation, stakeholder feedback)
- Learning from peer organisations

Output: 1-2 page evidence summary with ToC implications

Step 2: Assumption Testing

- Which assumptions confirmed by evidence?
- Which challenged or proven false?
- What new assumptions emerged?

Use Assumption Testing Matrix (B.1.2); update with new evidence and priorities

Step 3: Outcome Analysis

- Which outcomes achieved/stronger/weaker than expected?
- Any unexpected outcomes (positive or negative)?
- Are outcome timings accurate or need adjustment?
- Does ToC work equally for all participant groups?

Step 4: Implementation Learning

- Where does implementation diverge from ToC design?
- What adaptations have been made and why?
- Should ToC reflect current practice?

Step 5: Context Changes

- Policy environment shifts
- Economic/labour market conditions
- Partnership landscape changes
- Social/demographic factors

Step 6: Stakeholder Consultation

- Gather staff, participant, partner, management perspectives on ToC accuracy
- Where do stakeholder views diverge from ToC?

Method: Structured consultations specifically asking about ToC

Step 7: Revise ToC

- Update causal pathways
- Refine assumptions
- Adjust outcome timing
- Modify activities if needed
- Update visual and narrative

Step 8: Communicate

- Version control (date clearly, maintain change log, archive previous version)
- Share with all staff, partners, funders with explanation
- Update website, reports, training materials
- Adjust evaluation plan to match revised ToC

PART C: TRIGGERS FOR IMMEDIATE REVIEW (Don't wait for annual cycle)

- ! Unexpected outcomes (positive or negative) not in ToC
- ! Consistent participant feedback contradicting ToC logic
- ! Implementation significantly differs from ToC
- ! Key assumptions proven false
- ! Context fundamentally changed (major policy, economic shock, community trauma)
- ! Evaluation findings challenge causal pathways

PART D: REVISED TOC QUALITY CHECK

After updating, verify:

- Reflects current evidence (gaps acknowledged, sources cited)
- Captures implementation reality (staff recognise their work)
- Testable assumptions remain (not all certainties)
- Maintains usability (one page visual, 3-8 page narrative)
- Documents learning (change log explains what/why)
- Evaluation aligned (monitoring, data collection, reporting match revised ToC)

PART E: USING THE LIVING TOC

ToC is operational DNA, not compliance document.

Monthly/Quarterly: Reference in team meetings (service improvements, resource allocation, partnerships, training, risk management)

When unexpected results emerge: Ask "What does this tell us about our ToC?" If finding can't be traced to ToC, either finding is irrelevant OR ToC needs updating.

For new staff: Use as core training document (why we do this, how change happens, what we're learning, critical assumptions we're testing)

In reporting: Structure reports around ToC (progress on pathways, assumption testing results, context changes, refinements)

The acid test: Is ToC the first document you reach for in important programme decisions? If not, it needs to be more useful or accurate.

B.1.6 Service-Specific Theory of Change Examples

Purpose: Provide sector-specific ToC templates to guide development of programme-specific Theories of Change.

How to use: Review the sector closest to your programme; adapt the causal pathway structure, critical assumptions, and context factors to your specific context.

EXAMPLE 1: EDUCATION AND YOUTH SERVICES

Programme Context:

Target population: Young people (14-25 years) not in education, employment, or training (NEETs) or at risk of exclusion

Core service: Educational re-engagement, skills development, youth work support

Duration: 6-18 months

Key partners: Schools, colleges, youth services, employers, social services

Causal Pathway Structure:

Inputs: Youth workers, educational resources, safe spaces, funding, partnerships with education providers and employers, referral networks

Activities: One-to-one mentoring, group skills training, educational re-engagement support, work experience placements, personal development activities, crisis intervention, family engagement

Outputs: Young people engaged (target 100/year), 80% attendance, qualifications achieved, work placements completed, progression plans created

Short-term outcomes (0-6 months): Engagement with support established, confidence increased, basic skills improved, relationships with positive adults developed, crisis situations stabilised

Medium-term outcomes (6-18 months): Return to education/training (60%), employment secured (30%), qualifications achieved, social and life skills developed, family relationships improved

Long-term impacts (18+ months): Sustained education/employment, reduced offending, improved wellbeing, economic independence, positive adult roles

Critical Assumptions:

- **Youth-centred approach works:** Young people ARE experts in their own lives; given appropriate support, they CAN shape their futures (if false: youth-centred model won't work; revert to adult-led approaches)
- **Timing matters:** Early intervention (within 3 months of disengagement) prevents deeper exclusion (if false: 3-month guarantee unnecessary; timing flexibility acceptable)
- **Holistic approach necessary:** Addressing only education without life management factors (housing, family, mental health) produces limited outcomes (if false: can use education-only approach, don't need wraparound)
- **Accessible delivery essential:** Low-threshold, youth-friendly spaces are critical; high-quality services that are difficult to access won't reach those most in need (if false: traditional institutional settings adequate)
- **Relationships are mechanism:** Trusting relationships with youth workers are the primary change mechanism, not just activities delivered (if false: focus on programme content over relationship quality)



Context Factors:

- Economic: Youth unemployment rates, apprenticeship availability, benefit system rules for under-25s
- Policy: Education participation age, NEETs definitions, youth guarantee schemes, safeguarding frameworks
- Social: Youth stigma, gang involvement, family expectations, peer influences
- Geographic: Urban (more services, higher risks) vs. rural (isolation, transport barriers)

Testing Assumptions:

- Youth-centred: Compare outcomes between youth-led vs. adult-led programmes; gather youth voice on approach
- Timing: Track outcomes by time from disengagement to engagement (0-3 months vs. 3-6 vs. 6-12+)
- Holistic: Correlate addressing multiple needs with education/employment outcomes
- Accessibility: Measure engagement rates for low-threshold vs. traditional settings
- Relationships: Measure relationship quality; correlate with outcomes; compare continuity vs. multiple staff

EXAMPLE 2: EMPLOYMENT AND SKILLS PROGRAMMES

Programme Context:

- Target population: Unemployed or economically inactive adults (18-65 years)
- Core service: Skills assessment, vocational training, job matching, retention support
- Duration: 3-12 months
- Key partners: Employers, public employment services, training providers

Causal Pathway Structure:

Inputs: Funding, training facilities, staff (job coaches, trainers), employer partnerships, referral networks

Activities: Individual skills assessment, vocational training, work placements, job search support, employer engagement, in-work support

Outputs: 150 enrolled/year, 100% assessed, 80% complete training, 60 work placements, qualifications awarded, 70 job placements

Short-term outcomes (0-6 months): Skills acquired, confidence increased, job applications submitted, interviews attended, professional networks developed



Medium-term outcomes (6-18 months): Employment secured (70% target), 6-month retention (60%), earnings increased, career progression initiated

Long-term impacts (18+ months): Economic security, social inclusion, reduced poverty/benefit dependency, local economy contribution

Critical Assumptions:

- **Labour market demand exists:** Local labour market has opportunities in training areas provided (if false: training won't lead to employment; need to shift skill focus or target different markets)
- **Training matches employer needs:** Content matches current employer requirements and industry standards (if false: skills won't translate to employability; need curriculum redesign)
- **Dosage effects:** Participants completing 80%+ sessions will demonstrate measurable skills improvement (if false: focus on engagement quality not just completion)
- **Barriers can be addressed:** Participants can overcome non-skills barriers (health, housing, childcare, transport) through signposting and light-touch support (if false: need intensive wraparound or narrow eligibility)
- **Quality jobs accessible:** Jobs secured provide adequate wages and progression, not just entry to precarious work (if false: need employer engagement focused on job quality)

Context Factors:

- Economic: Local unemployment rate, sectoral growth/decline, benefit system rules
- Policy: Employment services reforms, welfare conditionality, qualification frameworks
- Social: Stigma against unemployed, employer discrimination, social networks
- Geographic: Urban/rural, transport infrastructure, employer concentration

Testing Assumptions:

- Labour market: Track placement rates; survey employers on skills needs; monitor labour market intelligence
- Training quality: Employer feedback; employment outcomes; wage levels achieved
- Dosage: Correlate attendance with skills assessment scores; pre-post tests
- Barriers: Track referral uptake and resolution; compare outcomes by barrier complexity
- Job quality: Track wages at 6/12/18 months; measure contract type, hours; progression indicators

EXAMPLE 3: MIGRATION AND REFUGEE INTEGRATION SERVICES

Programme Context:

- Target population: Refugees and asylum seekers newly arrived in host country
- Core service: Language learning, employment support, social integration, rights navigation
- Duration: 12-36 months (long integration journey)
- Key partners: Legal services, employment services, language providers, cultural organisations, schools, health services

Causal Pathway Structure:

Inputs: Multi-lingual staff and interpreters, language resources, employment support, legal services, cultural orientation materials, community spaces, emergency funding, partnership networks

Activities: Arrival support and orientation, host language classes (ESOL), employment preparation (skills recognition, job search), legal rights information and support, cultural integration activities, community connection and befriending, trauma-informed support, children's education support, health navigation

Outputs: 200 engaged/year, 200 hours language classes attended, employment support sessions, legal consultations, community events attended, peer support connections, children enrolled in schools, GP registrations completed

Short-term outcomes (0-12 months): Basic language skills (A1-A2), legal rights understood, basic needs met (housing, food, healthcare), safety established, initial social connections, trauma stabilisation, children settling in schools, cultural orientation

Medium-term outcomes (12-36 months): Functional language proficiency (B1-B2), employment secured (50% target within 24 months), housing stability, social networks developed (bonding + bridging capital), children integrated in schools, cultural navigation skills, health improved, qualification recognition progressing

Long-term impacts (36+ months): Economic integration and self-sufficiency, social integration and belonging, civic participation, intergenerational integration, contribution to host community, trauma recovery, family reunification (where applicable)

Critical Assumptions:

- **Language as foundation:** Host language proficiency (B1+) is foundational for employment, social integration, rights access (if false or only partially true: need multiple integration pathways; support co-ethnic economies; extensive translation/ interpretation)
- **Employment as primary pathway:** Employment is the main pathway to self-sufficiency, social connection, and integration (if false: need alternative pathways - education, volunteering, community leadership, caring; tackle discrimination; recognise unpaid care work)
- **Trauma recovery alongside integration:** Refugees can manage trauma whilst learning language, seeking work, integrating (if false for many: need trauma-specific services; phased approach - stabilisation → integration → advancement; trauma-informed practice essential)
- **Host community receptiveness:** Host communities are sufficiently welcoming for refugees to integrate socially and economically; integration is two-way (if false: community education essential; anti-discrimination work; cannot place all responsibility on refugees; challenge hostile narratives)
- **Qualification recognition enables appropriate employment:** Prior qualifications/ experience can be recognised, enabling skill-level employment (if false: lobby for better recognition; financial support for re-qualification; manage expectations; support alternative professional pathways)

Context Factors:

- Policy: Asylum policy, right to work, benefit access, housing policy (dispersal), immigration status uncertainty, citizenship pathways
- Economic: Labour market conditions, skills demand, qualification recognition barriers, discrimination in employment/housing, living costs
- Social: Public attitudes to asylum, media portrayal, community cohesion, existing diaspora communities, hate crime levels, civil society strength
- Geographic: Urban (services, opportunities, diversity) vs. rural (isolation, limited services), dispersal patterns, transport, housing availability

Testing Assumptions:

- Language foundation: Correlate language proficiency (A1/A2/B1/B2) with employment, social integration, rights access; identify non-language pathways
- Employment centrality: Compare integration indicators for employed vs. unemployed; identify alternative pathways; measure job quality not just employment rate
- Trauma/integration: Track relationship between trauma indicators and integration outcomes (ethical/methodological challenges); monitor mental health access

- Community receptiveness: Measure discrimination experiences; employment/housing access patterns; hate crime data; community cohesion indicators
- Qualification recognition: Track recognition rates and barriers; measure skills match in employment; cost-benefit of re-qualification support

Special Evaluation Considerations:

- Long timeframe (36+ months minimum for meaningful integration outcomes)
- High mobility (refugees move, difficult follow-up)
- Language barriers in data collection (interpreted surveys, translated materials essential)
- Ethical issues (measuring trauma, trust with researchers who may be confused with authorities)
- Diverse populations (Syrians, Afghans, Eritreans have different contexts)
- Attribution challenges (integration affected by policy, discrimination, economic conditions far beyond programme)
- Disaggregation essential (nationality, gender, age, family composition, education, arrival route, asylum status)

B.2: OUTCOME MONITORING TEMPLATES AND TOOLS

B.2.1 Outcome Measurement Planning Template

Measurement Specifications

Template 5 – Outcome Monitoring Statement Template

Element	Specification
Measurement tool	[Name of scale/instrument or description of bespoke measure]
Validated?	Yes/No [If yes, cite source]
Response format	[e.g., 1-5 scale, yes/no, 0-10 rating, frequency count]
Data collection method	[Self-completed, staff-administered, administrative data]
Baseline timing	[When in service journey]
Follow-up timing	[Specify all follow-up points]
Target sample	[All participants or sampling strategy]
Responsible staff	[Role]
Data storage	[System/location]
Quality checks	[How monitored]

[Repeat for each outcome]

Template 6 – Outcome Monitoring Quality Assurance Plan

Quality Dimension	Target	How Monitored	Frequency	Action if Target Missed
Baseline completion rate	e.g. ≥80%	Database report	Monthly	[Corrective action]
Follow-up completion rate	e.g. ≥60%	Database report	Monthly	[Corrective action]
Data completeness	≥95% fields completed	Random audit	Monthly	[Corrective action]
Data consistency	<5% inconsistencies	Logic checks	Weekly	[Corrective action]
Staff compliance	All staff trained	Training register	Quarterly	[Corrective action]

Template 7 – Outcome Monitoring Analysis and Reporting Schedule

Reporting Period	Analysis Required	Report Format	Audience	Due Date
Monthly	Basic descriptive statistics	1-page summary	Team meeting	[Day of month]
Quarterly	Full analysis with disaggregation	3-5 page report	Management, staff	[Date]
Annual	Comprehensive analysis, trends	Formal report	Funders, governance	[Date]

B.2.2 Sample Baseline and Follow-Up Data Collection Forms

PART A: PARTICIPANT INTAKE FORM (Collected Once at Enrolment)

[PROGRAMME NAME] - PARTICIPANT INTAKE FORM

Programme ID: [Unique identifier - assign sequentially]

Enrolment Date: [DD/MM/YYYY]

Staff Member: [Name]

Section 1: Core Demographics

Date of Birth: [DD/MM/YYYY]

Age: [Auto-calculate or record]

Gender:

Male Female Non-binary Prefer not to say Prefer to self-describe: _____

Postcode: [First 4-5 characters for geographic analysis]

Ethnicity: [Use categories appropriate for national context]

[Category 1]

[Category 2]

[Category 3]

Prefer not to say

Highest Education Level:

No formal qualifications

Lower secondary (ISCED 2)

Upper secondary (ISCED 3)

Post-secondary non-tertiary/Vocational (ISCED 4)

Tertiary - Bachelor's or equivalent (ISCED 6)

Tertiary - Master's or equivalent (ISCED 7)

Prefer not to say

Section 2: Programme-Specific Baseline

[ADAPT THIS SECTION TO YOUR SERVICE TYPE - Examples provided below]

EXAMPLE FOR EMPLOYMENT PROGRAMMES:

Current Employment Status:

Employed full-time

Employed part-time

Self-employed

Unemployed and actively seeking work

Unemployed and not seeking work



- Student
- Unable to work (health/disability)

If unemployed, duration: _____ months

Main support needs (select up to 3):

- Skills training
- Job search support
- Work experience
- Confidence building
- Childcare support
- Transport support
- Language support
- Other: _____

EXAMPLE FOR EDUCATION/YOUTH PROGRAMMES:

Current Status:

- In education/training
- Not in education, employment, or training (NEET)
- Employed but seeking education
- Other: _____

If NEET, duration: _____ months

Last education attended: [Name of school/college/provider]

Date left: [MM/YYYY]

Reason for leaving (if applicable):

- Completed course
- Exclusion
- Financial reasons
- Caring responsibilities
- Mental health
- Disengagement
- Other: _____

Main support needs (select up to 3):

- Re-engagement with learning
- Qualification support
- Life skills
- Mentoring
- Financial support

- Mental health support
- Other: _____

EXAMPLE FOR MIGRATION/REFUGEE SERVICES:

Arrival in [Country]: [MM/YYYY]

Time in country: _____ months

Immigration Status:

- Refugee status granted
- Asylum seeker (decision pending)
- Humanitarian protection
- Family reunion
- Other: _____
- Prefer not to say

Country of Origin: _____ [optional]

Languages Spoken:

First language: _____

Host language proficiency (self-rated): None Basic (A1-A2) Intermediate (B1-B2)
 Advanced (C1-C2)

Main support needs (select up to 3):

- Language learning
- Employment support
- Housing support
- Legal/rights information
- Healthcare navigation
- Children's education support
- Social connection
- Other: _____

Section 3: Referral Information

Referral Source:

- Self-referral
- Public employment/social services
- Health services
- Another NGO: _____
- Friend/family recommendation
- Other: _____

**Previous participation in similar programmes:** No Yes → Programme name: _____**PART B: OUTCOME MEASUREMENT FORM (Collected at Baseline, Exit, Follow-up)****[PROGRAMME NAME] - OUTCOME MEASUREMENT****Programme ID:** [Match to intake form]**Assessment Type:** Baseline Exit Follow-up (____ months post-exit)**Assessment Date:** [DD/MM/YYYY]**Instructions**

This form measures progress on programme outcomes. Select 3-5 priority outcomes from your Theory of Change and include appropriate measurement tools below. Use identical questions at all time points (baseline, exit, follow-up) to enable comparison.

OUTCOME 1: [Name - e.g., Employment Confidence, Wellbeing, Language Proficiency]**[INSERT YOUR CHOSEN MEASURE - Examples provided]****EXAMPLE: Employment Confidence (4-item scale)**

How confident do you feel about:

1. Finding suitable job opportunities?
1 (Not at all confident) --- 2 --- 3 --- 4 --- 5 (Very confident)
2. Completing job applications successfully?
1 --- 2 --- 3 --- 4 --- 5
3. Performing well in job interviews?
1 --- 2 --- 3 --- 4 --- 5
4. Securing and maintaining suitable employment?
1 --- 2 --- 3 --- 4 --- 5

Total Score: _____ / 20



EXAMPLE: General Wellbeing (ONS4 Questions - UK standard)

- 5. Overall, how satisfied are you with your life nowadays?
0 (Not at all) ----- 10 (Completely)
- 6. Overall, to what extent do you feel things you do in your life are worthwhile?
0 (Not at all) ----- 10 (Completely)
- 7. Overall, how happy did you feel yesterday?
0 (Not at all) ----- 10 (Completely)
- 8. Overall, how anxious did you feel yesterday?
0 (Not at all) ----- 10 (Completely)

EXAMPLE: Language Proficiency (CEFR Self-Assessment)

Speaking: I can...

- A1: Use simple phrases about myself and familiar matters
- A2: Communicate in simple routine tasks
- B1: Deal with most situations in areas where language is spoken
- B2: Interact with fluency and spontaneity with native speakers
- C1: Express myself fluently and spontaneously
- C2: Express myself with precision and ease

OUTCOME 2: [Name]

[INSERT YOUR CHOSEN MEASURE]

[Space for questions]

Total Score: ____ / ____

OUTCOME 3: [Name]

[INSERT YOUR CHOSEN MEASURE]

[Space for questions]

Total Score: ____ / ____

OPTIONAL: Outcome 4 & 5

[Only if resources permit - remember: better to measure fewer outcomes well than many outcomes poorly]

Brief Qualitative Feedback (Exit and Follow-up only)

What has changed most for you since starting this programme?

[Open text - 2-3 sentences]

What helped you most?

[Open text - 1-2 sentences]

What barriers remain? [Follow-up only]

[Open text - 1-2 sentences]

GUIDANCE NOTES

Purpose of Two Forms:

- **Intake Form:** Collected ONCE at enrolment; captures demographics, baseline status, support needs
- **Outcome Form:** Collected MULTIPLE TIMES (baseline, exit, follow-up); tracks change on priority outcomes

Benefits of Separation:

- Outcome form is brief (10-15 minutes) - higher completion rates
- Demographics don't need repeating at each measurement point
- Clearer for staff: administrative data vs. outcome data

Timing:

- **Intake Form:** At enrolment (before service starts)
- **Outcome Form - Baseline:** Within first 2 weeks of service
- **Outcome Form - Exit:** At programme completion or after defined period (e.g., 6 months)
- **Outcome Form - Follow-up:** 3-6 months post-exit (if resources permit)

Customisation:

- Adapt Section 2 of Intake Form to your service type
- Select 3-5 priority outcomes from your Theory of Change for Outcome Form
- Use validated tools from Annex B.2.5 where available
- Keep total Outcome Form under 15 questions for feasibility

Data Linking:

- Use same Programme ID on both forms to link participant's intake and outcome data
- Store securely; never use names in analysis files, only IDs

B.2.3 Simple Analysis Spreadsheet Template

Template 8 – Outcome Monitoring Participant Register Template

ID	Baseline Date	Exit Date	Follow-up Date 1	Follow-up Date 2	Status	Cohort	Notes
1	01/01/2024	01/04/2024	01/07/2024		Completed	2024-Q1	
2	05/01/2024				Active	2024-Q1	

Template 9 – Outcome Monitoring Outcome Data Template

ID	Gender	Age	Outcome 1 Baseline	Outcome 1 Exit	Outcome 1 Change	Improved?
1	F	28	8/20	16/20	8	Y
2	M	35	12/20	15/20	3	Y
3	F	42	15/20	15/20	0	N

Formulas:

- Change = [Exit score] - [Baseline score]
- Improved? = IF([Change]>0, "Y", "N")

Sheet 3: Summary Analysis

Completion Rates:

- Total participants enrolled: [COUNT]
- Baseline completed: [COUNT] ([PERCENTAGE]%)
- Exit completed: [COUNT] ([PERCENTAGE]%)
- Follow-up completed: [COUNT] ([PERCENTAGE]%)



Outcome 1: [Name]

- Mean baseline score: [AVERAGE]
- Mean exit score: [AVERAGE]
- Mean change: [AVERAGE]
- % improved: [PERCENTAGE]%
- % stayed same: [PERCENTAGE]%
- % declined: [PERCENTAGE]%

Outcome 2: [Name] [Same structure]

Disaggregated Analysis:

Disaggregated Analysis By Gender

Gender	n	Mean Baseline	Mean Exit	Mean Change	% Improved
Female					
Male					

Disaggregated Analysis By Age group

Age Group	n	Mean Baseline	Mean Exit	Mean Change	% Improved
18-25					
26-40					
41-60					
60					

B.2.4 Outcome Monitoring Report Template

[PROGRAMME NAME] - OUTCOME MONITORING REPORT

Reporting Period: [e.g., January-June 2025 OR Q1-Q2 2025 OR Annual 2024-25]

Report Date: [DD/MM/YYYY]

Prepared by: [Name, Role]

1. EXECUTIVE SUMMARY

Key Finding: [One sentence on most important outcome or pattern]

Main Concern: [One sentence on biggest challenge or disappointing result]

Priority Action: [One sentence on what will change next quarter based on findings]

2. PARTICIPATION THIS PERIOD

Template 10 – Outcome Monitoring Participation Reporting

Metric	Target	Actual	Status
New participants enrolled	[N]	[N]	<input type="checkbox"/> On track <input type="checkbox"/> Below target
Active participants	[N]	[N]	<input type="checkbox"/> On track <input type="checkbox"/> Below target
Completed programme	[N]	[N]	<input type="checkbox"/> On track <input type="checkbox"/> Below target
Dropped out	-	[N]	[%]
Data completion rates:			
Baseline forms completed	>90%	[%]	<input type="checkbox"/> Adequate <input type="checkbox"/> Needs improvement
Exit forms completed	>70%	[%]	<input type="checkbox"/> Adequate <input type="checkbox"/> Needs improvement

Demographics of participants this period:

- Gender: ___% Female, ___% Male, ___% Other/Prefer not to say
- Age: ___% 18-25, ___% 26-40, ___% 41-60, ___% 60+
- [Other relevant demographic breakdowns]

Notes on participation: [2-3 sentences on any notable patterns - e.g., higher dropout than usual, different demographic mix, capacity issues]

3. OUTCOME RESULTS

Outcome 1: [Name, e.g., Employment Confidence]

Measure	Baseline Mean	Exit Mean	Mean Change	% Improved
[Outcome 1]	[Score]	[Score]	[+/-]	[%]

Target met? Yes No

Interpretation: [2-3 sentences - is this change meaningful? How does it compare to targets or previous quarters?]

Outcome 2: [Name, e.g., Wellbeing]

Measure	Baseline Mean	Exit Mean	Mean Change	% Improved
[Outcome 2]	[Score]	[Score]	[+/-]	[%]

Target met? Yes No

Interpretation: [2-3 sentences]

Outcome 3: [Name]

Measure	Baseline Mean	Exit Mean	Mean Change	% Improved
[Outcome 3]	[Score]	[Score]	[+/-]	[%]

Target met? Yes No

Interpretation: [2-3 sentences]

Disaggregated Results: Who Benefits Most?

[Choose ONE key demographic or characteristic to examine this period - rotate focus each period if reporting frequently]

This period's focus: [e.g., Gender / Age group / Duration of unemployment / Baseline severity]

Group	N	Mean Change on [Primary Outcome]	% Improved
[Group A - e.g., Female]	[N]	[+/-]	[%]
[Group B - e.g., Male]	[N]	[+/-]	[%]

Key finding: [2 sentences - are outcomes similar across groups or are there disparities? Why might this be?]

4. WHAT WE LEARNED

Patterns and Insights

[2-3 paragraphs covering:]

- What's working well? (e.g., "Participants who attended 80%+ sessions showed twice the improvement...")
- What's not working as expected? (e.g., "Younger participants (18-25) had higher dropout rates...")
- Any surprises or unexpected findings? (e.g., "Wellbeing improved even for those who didn't secure employment...")
- What do participant comments tell us? (themes from qualitative feedback)

Illustrative Case

[OPTIONAL: 3-4 sentences describing one participant's journey that illustrates typical change or an important learning point - use anonymised details]

Example: "Maria, 34, enrolled with very low confidence (score 6/20). Despite facing childcare challenges that caused her to miss several sessions, she engaged consistently with job search support. By exit, her confidence had increased to 16/20, and she secured part-time employment. This illustrates how targeted support can yield outcomes even with imperfect attendance."

5. DATA QUALITY AND LIMITATIONS

Data quality this period:

- Good - minimal issues
- Acceptable - some issues managed
- Poor - significant problems

Issues identified and how addressed:

- [Issue 1 - e.g., "Three baseline forms completed late; reminded staff of timing protocol"]
- [Issue 2 - e.g., "Two exit surveys had missing items; followed up for completion"]

Limitations to consider when interpreting results:

- [Limitation 1 - e.g., "Small sample size this quarter (n=15) limits reliability"]
- [Limitation 2 - e.g., "Unable to conduct follow-up for 4 participants who moved areas"]
- [Limitation 3 - e.g., "Self-reported outcomes; no objective verification"]

6. NEXT PERIOD PRIORITIES

Based on this period's findings, our priorities for [Next Period - e.g., next 6 months, next year] are:

1. **[Priority 1]** - [Action to address main concern or build on success]
Responsible: [Name/Team] | Deadline: [Date]
2. **[Priority 2]** - [Action]
Responsible: [Name/Team] | Deadline: [Date]
3. **[Priority 3]** - [Action]
Responsible: [Name/Team] | Deadline: [Date]

Data collection focus:

- [e.g., "Improve exit form completion rate from 65% to >75%"]
- [e.g., "Add brief question about attendance barriers to identify patterns"]

Target length: 2-3 pages maximum - concise is better than comprehensive

Audience: Primarily internal (staff, management) but can be adapted for funders

Key principles:

- **Honest:** Report disappointing results as well as successes; learning comes from both
- **Action-oriented:** Every finding should prompt a question: "So what will we do differently?"
- **Accessible:** Avoid statistical jargon; present numbers simply
- **Balanced:** Include quantitative data AND qualitative insights/case examples

Tips for effective reporting:

- Use visuals where helpful (simple bar charts, line graphs showing trends over reporting periods)
- Compare to previous periods to show trends (improving/stable/declining)
- Highlight one "deep dive" each period (e.g., Period 1 focus on gender, Period 2 focus on age)
- Discuss in team meeting: 30-45 minutes for presentation and discussion of findings and priorities
- Archive reports to track learning over time

What to do if outcomes are disappointing:

- Don't ignore or downplay - investigate why
- Possible explanations: Unrealistic targets? Wrong outcomes measured? Implementation fidelity issues? External factors (e.g., pandemic, policy changes)? Wrong participants recruited? Insufficient dosage?
- Use findings to adjust programme, not just report
- Remember: Monitoring is for learning and improvement, not just accountability

Annual synthesis:

- If reporting more frequently than annually, produce brief annual summary highlighting: overall trends, major learnings, programme adaptations made based on evidence

B.2.5 Validated Measurement Tools by Service Type

EMPLOYMENT SERVICES

Work Readiness Scale

- 8-item scale measuring job search self-efficacy
- 5-point Likert responses
- Public domain
- Source: Smith & Betz (2000). Development and validation of a scale of perceived social self-efficacy. *Journal of Career Assessment*, 8(3), 283-301.

Career Decision Self-Efficacy Scale - Short Form

- 25-item scale
- 5-point confidence ratings
- Requires permission from authors
- Source: Betz & Taylor (2012)

MENTAL HEALTH SERVICES

Patient Health Questionnaire (PHQ-9)

- 9-item depression screening tool
- Widely used in clinical and research settings
- Public domain
- Source: Kroenke, Spitzer & Williams (2001). *Journal of General Internal Medicine*, 16(9), 606-613.

Generalised Anxiety Disorder Scale (GAD-7)

- 7-item anxiety screening tool
- 4-point frequency scale
- Public domain
- Companion to PHQ-9

Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS)

- 14-item (or 7-item short version) wellbeing scale
- Validated across populations
- Free for non-commercial use
- Source: NHS Health Scotland, University of Warwick, University of Edinburgh

HOUSING SERVICES

Homelessness Outcomes Star

- 10-dimension outcome measurement tool
- Journey of change approach (1-5 ladder)
- Requires licence (modest cost)
- Source: Triangle Consulting Social Enterprise

Housing Stability Scale

- Measures housing security perceptions
- Public domain alternative to Outcomes Star
- Can be adapted to local context

Wellbeing (General)

ONS4 Wellbeing Questions

- 4-item national wellbeing measures (UK)

- Life satisfaction, happiness, anxiety, worthwhile
- 0-10 scales
- Public domain, enables national benchmarking
- Source: UK Office for National Statistics

WHO-5 Wellbeing Index

- 5-item general wellbeing measure
- 6-point frequency scale
- Public domain
- Available in 30+ languages
- Source: World Health Organisation

Social Connections

Lubben Social Network Scale (LSNS)

- 6-item (short) or 18-item (full) social isolation measure
- 6-point frequency responses
- Public domain
- Source: Lubben et al. (2006). *The Gerontologist*, 46(4), 503-513.

UCLA Loneliness Scale

- 20-item or 3-item (ULS-3) short version
- Measures subjective loneliness
- Public domain
- Source: Russell (1996)

Service Satisfaction (Generic)

Client Satisfaction Questionnaire (CSQ-8)

- 8-item general satisfaction measure
- 4-point scales
- Public domain
- Source: Larsen et al. (1979). *Evaluation and Program Planning*, 2(3), 197-207.

Net Promoter Score (NPS)

- Single-item recommendation likelihood (0-10)
- Widely used, simple
- Public domain concept
- Follow-up: "Why did you give this score?"

EDUCATION & YOUTH SERVICES

Academic Self-Efficacy Scale (ASES)

- 8-item scale measuring confidence in academic tasks
- 5-point Likert scale
- Public domain
- Relevant for: Re-engagement programmes, adult learning, skills training

Rosenberg Self-Esteem Scale

- 10-item scale, widely validated with youth
- 4-point Likert
- Public domain
- Relevant for: Youth development, education re-engagement

School/Programme Engagement Scale

- 6-item behavioural engagement sub-scale
- Attendance + participation + effort
- Adaptable to non-school education settings

Educational Aspiration Index

- Brief (4-item) measure of educational goals and expectations
- Relevant for: NEETs programmes, adult education

Basic Skills Assessment

- Literacy: Use national frameworks (e.g., UK Functional Skills descriptors)
- Numeracy: Use national frameworks
- Digital skills: Use Digital Skills Framework levels

MIGRATION & REFUGEE INTEGRATION

Host Language Proficiency

- Use CEFR (Common European Framework) self-assessment grid (A1-C2)
- Public domain, standardised across EU
- 6 levels with clear descriptors for speaking, listening, reading, writing

Integration Progress Scale

- Adapted from various national integration monitoring tools
- Domains: Language, Employment, Housing, Social Connection, Cultural Navigation

- 5-point self-rated scales per domain

Acculturation Scale (Brief)

- Berry's 8-item bidimensional acculturation scale
- Measures heritage culture maintenance + host culture adoption
- Public domain

Perceived Discrimination Scale

- Brief 5-item everyday discrimination scale
- Important context for integration outcomes
- Public domain

Social Capital Scale for Migrants

- Bonding capital (connections within own community) vs. Bridging capital (connections with host community)
- 8-item scale

Employment/Education Goal Progress

- Simple self-rated: "To what extent have you progressed toward your employment/education goals since arrival?" (0-10 scale)
- Qualitative follow-up: "What barriers remain?"

B.2.6 Data Quality Checklist

Monthly Review

- Completeness: baseline and follow-up form completion rates; patterns in missing data
- Consistency: logic checks, impossible values, duplicates, free-text errors
- Accuracy: spot-check 5–10 forms against source documents; retraining if error rate >5%
- Timeliness: lag between service contact and data entry; address barriers
- Follow-up tracking: approaching deadlines, contact attempts, tracking system updates
- Documentation: log issues, corrective actions, systematic problems; share with team

Quarterly Audit

Comprehensive review: database anomalies, quality trends, comparison to targets
 Staff compliance: identify training needs; address non-compliance Process review:
 procedures, form usability, timing System functionality: database performance,
 backup, security Reporting: data quality report for management; update procedures;
 set improvement goals

RESPONDING TO DATA QUALITY PROBLEMS

Error rate	Response
<5%	Individual feedback; monitor for patterns
5–15%	Refresher training; review procedures; increase spot checks
>15%	Immediate procedural review; one-to-one coaching; simplify forms if needed
>20% missing forms	Investigate root cause (burden, timing, resistance) before assuming non-compliance

Remember: Data quality problems are typically system failures, not individual ones. Address procedures, forms, and resourcing before attributing problems to staff.

B.3: STAKEHOLDER FEEDBACK TEMPLATES AND TOOLS

B.3.1 Stakeholder Mapping Template

Purpose: Identify all stakeholders, prioritise feedback collection, and plan proportionate engagement.

STAKEHOLDER MAPPING MATRIX

Template 11 – Stakeholder Mapping Matrix

Stakeholder Group	Unique Insights They Offer	Priority (H/M/L)	Feedback Method	Frequency	Notes
PARTICIPANTS/SERVICE USERS					
[e.g., Programme participants]	Service experience, outcomes, barriers, satisfaction	H	Satisfaction survey	Exit + 6-month follow-up	Target >60% response
[e.g., Family members/carers]	Wider impacts, support needs, family perspective	M	Brief survey or focus group	Annual	If relevant to service
[e.g., Drop-outs/non-completers]	Why disengagement, barriers, unmet needs	M	Exit interview (phone)	As occur	Often hardest to reach
STAFF					
[e.g., Front-line delivery staff]	Implementation challenges, participant feedback, improvement ideas	H	Team meetings + annual pulse survey	Ongoing + annual	5-question pulse check
[e.g., Management]	Strategic issues, resource constraints, partnership opportunities	M	Regular meetings	Quarterly	Often informal
PARTNERS & REFERRERS					
[e.g., Referral agencies]	Referral appropriateness, communication, outcomes from their perspective	H	Brief feedback form	Annual	Key for continuation

[e.g., Delivery partners]	Joint working effectiveness, coordination, gaps/overlaps	M	Partnership review meeting	6-monthly	Focus on collaboration
[e.g., Specialist services (health, housing, legal)]	Integration effectiveness, unmet needs, system barriers	L	Informal consultation	As needed	Ad hoc basis
FUNDERS & COMMISSIONERS					
[e.g., Primary funder]	Alignment with priorities, value for money, governance	H	Formal reporting + meetings	Quarterly	Contractual requirement
[e.g., Grant funders]	Impact evidence, sustainability, learning	M	Annual reports + case studies	Annual	Varied requirements
COMMUNITY & OTHER					
[e.g., Local community representatives]	Community impacts, reputation, local integration	L	Consultation as needed	Ad hoc	If community allows
[e.g., Policy makers]	Policy relevance, scalability, system change needs	L	Formal submissions	As opportunities arise	Strategic engagement

GUIDANCE FOR COMPLETING THE MATRIX

Column 1: Stakeholder Group

- List ALL groups who interact with or are affected by your programme
- Be specific (not just "partners" but which partners)
- Include those who might have criticism or concerns

Column 2: Unique Insights They Offer

- What can THIS group tell you that others can't?
- What perspective or information is specific to their role/relationship?
- Be clear about what you want to learn from them

Column 3: Priority (High/Medium/Low)

HIGH priority:

- Central to service delivery and quality
- Feedback is actionable and affects decisions
- Legally/contractually required (e.g., funders)
- Example: Participants, key referrers, primary funder

MEDIUM priority:

- Important but consult less frequently
- Useful intelligence but not always actionable
- Nice to have but not essential
- Example: Delivery partners, specialist services, some staff groups

LOW priority:

- Peripheral to core service
- Limited capacity to act on feedback
- Consult opportunistically not systematically
- Example: General community, policy makers, media

Column 4: Feedback Method

- Match method to stakeholder group and what you want to learn
- Options: Survey, interview, focus group, meeting discussion, informal conversation, observation
- Consider stakeholder preferences and accessibility

Column 5: Frequency

- How often will you seek feedback from this group?
- Options: Exit/completion, 3-monthly, 6-monthly, annual, ad hoc, ongoing
- Balance learning needs with consultation fatigue risk

Column 6: Notes

- Response rate targets
- Specific issues to explore
- Barriers to engagement
- Links to other processes
- Who is responsible for collection

B.3.2 Participant Satisfaction Survey (Generic Template)

[PROGRAMME NAME] - PARTICIPANT FEEDBACK SURVEY

Thank you for taking the time to share your views. Your feedback helps us improve our services. This survey is anonymous unless you choose to provide your name.

Completion date: [DD/MM/YYYY]

How you're completing this: Paper form Online Telephone Face-to-face with staff

SECTION 1: OVERALL SATISFACTION

1. Overall, how satisfied are you with [programme name]?

- Very satisfied
 Satisfied
 Neither satisfied nor dissatisfied
 Dissatisfied
 Very dissatisfied

2. How likely are you to recommend [programme name] to someone in a similar situation?

0 (Not at all likely) --- 1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10 (Extremely likely)

Why did you give this score? [Open text]

SECTION 2: ACCESS AND ENGAGEMENT

3. How easy was it to access our service?

Finding information about the service:

- Very easy Easy Neither easy nor difficult Difficult Very difficult Not applicable

Referral/application process:

- Very easy Easy Neither easy nor difficult Difficult Very difficult

Waiting time from referral to first contact:

- Much shorter than expected About right Longer than expected

Location/accessibility of service:

- Very convenient Convenient Neither convenient nor inconvenient Inconvenient
 Very inconvenient

4. Were there any barriers to participating fully? (Select all that apply)

- No barriers



- Transport difficulties
- Timing of sessions
- Childcare
- Language
- Disability access
- Cost
- Other: _____

SECTION 3: SERVICE DELIVERY

5. Please rate the following aspects of the service:

Staff were respectful and supportive:

1 (Strongly disagree) --- 2 --- 3 --- 4 --- 5 (Strongly agree)

Staff listened to my needs and concerns:

1 --- 2 --- 3 --- 4 --- 5

I was involved in decisions about my support:

1 --- 2 --- 3 --- 4 --- 5

The service was relevant to my needs:

1 --- 2 --- 3 --- 4 --- 5

Communication was clear and timely:

1 --- 2 --- 3 --- 4 --- 5

SECTION 4: OUTCOMES AND HELPFULNESS

6. Has [programme name] helped you?

Very much Quite a bit Somewhat A little Not at all

7. In what ways has the service been helpful? [Open text - 2-3 sentences]

8. What could be improved? [Open text - 2-3 sentences]

SECTION 5: ABOUT YOU (OPTIONAL)

These questions help us understand if the service works equally well for everyone. All answers are confidential.

Age:

Under 18 18-25 26-40 41-60 Over 60 Prefer not to say

Gender:

Male Female Non-binary Prefer not to say Prefer to self-describe: _____

Ethnicity: [Use categories appropriate for national context]

- [Category 1]
- [Category 2]
- [Category 3]
- Prefer not to say

Programme status:

- Yes, completed as planned
- Ongoing
- Left early
- Prefer not to say

Would you be willing to participate in a short interview to discuss your experience in more detail?

- Yes No

If yes, please provide contact details:

Name: _____

Email/Phone: _____

Thank you for your feedback. Your views are very important to us and will help improve services for others.

B.3.3 Focus Group Facilitation Protocol

SETUP (DO THIS BEFORE THE SESSION)

Write your key question: "Why do people drop out?" NOT "Let's hear from participants"

Invite enough people: Expect around 30% no-shows

Keep groups similar: All completers OR all drop-outs (not mixed)

Book: Accessible room, 90-120 minutes, refreshments

Bring: Consent forms, name cards, audio recorder

Write 5-8 questions:

1. Easy warm-up: "What brought you here?"
- 2-4. Your key topics with probes ready
2. Closing: "One thing to change?"

RUNNING THE SESSION

Start (10 min):

1. Welcome. Explain purpose: "We want to understand X to improve Y"
2. Ground rules: "Everyone's view matters. Speak one at a time. Confidential. No wrong answers."
3. Get consent. Record if they agree (tell them why)
4. Name cards, first names only

Warm-up (5 min): Ask something easy everyone can answer: "One word for your experience?" or "Why did you join?"

Main discussion (60-80 min):

Ask your key questions. Use this structure for each:

- **Ask openly:** "Tell me about staff support" NOT "Staff were great, right?"
- **Let them talk:** Wait 10-15 seconds of silence before prompting
- **Probe:** "Example?" "Anyone different?" "Why do you think that?"
- **Redirect if needed:** "That's interesting - let's come back to X"

Close (5 min):

1. Summarise what you heard in 3 bullets
2. Ask: "Missed anything?"
3. Tell them what happens next
4. Give vouchers, thank them

AFTER THE SESSION

Immediately (same day):

- Write 5 bullet points: main themes you heard
- Note any strong disagreements or surprises
- Note group dynamics that affected discussion

Within 48 hours:

- Listen to recording OR read notes
- Write 1-page summary:

- What did most people say? (consensus themes)
- Where did people disagree? (divergent views)
- What surprised you? (unexpected insights)
- 3-5 quotes that capture key points (anonymise names)

Integration:

- Compare to your survey data: Do focus groups explain the "why" behind numbers?
- Look for what surveys couldn't tell you: emotional responses, contextual factors, unintended impacts

Reporting:

- Write THEMES not individual stories: "Participants valued flexible timing" NOT "Sarah said timing was good"
- Use quotes to illustrate: "One participant said: *'The evening sessions meant I could work during the day'*"
- Be honest about group composition: "6 women, ages 25-45, all completers - drop-outs' views not represented"
- Don't over-claim: 6 people ≠ everyone; this is depth not breadth

Remember: Focus groups tell you WHY (surveys tell you WHAT). You're listening for depth, context, emotion, surprise - not counting responses.

B.3.4 "You Said, We Did" Communication Template

[ORGANISATION NAME] - FEEDBACK UPDATE

Period: [Quarter/Year]

Date: [Date]

Thank you to everyone who shared feedback about our services. We've carefully reviewed everything you told us. Here's what we learned and what we're doing about it.

WHAT YOU SAID

Issue 1: [Brief description]

What we heard: "[Representative quote or summary]"

How many people raised this: [e.g., "8 participants mentioned this" or "This came up in both staff survey and participant feedback"]

WHAT WE'VE DONE



Actions taken:

- [Specific action 1]
- [Specific action 2]
- [Specific action 3]

Impact so far: [Brief description of results]

What's still to come: [Any ongoing work]

WHAT YOU SAID

 **Issue 2: [Description]**

What we heard: "[Quote/summary]"

How many people raised this: [n]

 **WHAT WE'VE DONE**

[Same structure]

WHAT YOU SAID

 **Issue 3: [Description]**

What we heard: "[Quote/summary]"

How many people raised this: [n]

 **WHAT WE'RE WORKING ON**

Why we haven't acted yet: [Honest explanation - e.g., "This requires additional funding which we're currently seeking" or "We need to consult with partners before making changes"]

What we're doing:

- [Step 1]
- [Step 2]

Expected timeline: [When you expect to resolve this]

WHAT WE CAN'T CHANGE (AND WHY)

 **Issue 4: [Description]**



Why we can't address this: [Honest explanation - e.g., "This is determined by funder requirements" or "This would require resources beyond our budget" or "This conflicts with feedback from other stakeholders"]

What we can do instead: [Alternative approach if possible]

WHAT YOU TOLD US WE'RE DOING WELL

We're pleased you value:

- [Strength 1] - "We'll keep doing this"
- [Strength 2] - "This is important to us too"
- [Strength 3] - "Thank you for recognising this"

WHAT'S NEXT

Our priorities for [next period]:

1. [Priority 1]
2. [Priority 2]
3. [Priority 3]

When we'll ask for feedback again: [Date/timing]

How to give us feedback anytime:

- Suggestion box [location]
- Email: [address]
- Speak to any staff member
- [Other channels]

Thank you again for helping us improve. Your voice matters.

[Organisation contact details]

B.3.5 Feedback Analysis Framework

STEP 1: ORGANISE YOUR DATA

Ratings data:

- One spreadsheet row per respondent
- Columns: ID, demographics, all rating questions, date
- Calculate: n, mean, % positive (4-5 on 5-point scale)

Comments:

- One document with all comments
- Label each: ID, stakeholder group, date
- Read everything once before coding

Multiple groups (participants, staff, partners):

- Analyse separately first, then compare

STEP 2: ANALYSE RATING SCALES**Calculate summary:**

Question	N	Mean	% Positive
Overall satisfaction	45	4.2/5	78%
Staff respectful	45	4.6/5	91%
Service relevant	45	3.8/5	64%

Read the numbers:

- Mean >4.0 = Strong
- Mean 3.5-4.0 = Acceptable
- Mean <3.5 = Concern
- High negative ratings = Red flag

Find patterns:

- Highest scores = Your strengths
- Lowest scores = Your priorities
- Big % negative = Urgent attention needed

Compare groups (if n>30):

By gender	N	Mean
Female	28	4.3
Male	17	3.9

Does satisfaction vary by age, gender, etc? If yes, investigate why.



Track over time:

Period	Mean	Change
2024 H1	4.1	-
2024 H2	4.2	+0.1 (improving)

Trends matter more than absolute scores.

STEP 3: ANALYSE COMMENTS

Read once: Get overall sense (positive? critical? mixed?)

Code themes:

Theme	Count	Example Quotes
Staff quality	7	"Staff really listened"
Access barriers	3	"Hard to get to with buses"
Outcomes	5	"I feel much more confident now"

Look for:

- **Consensus** (many say same thing) = Common theme
- **Divergence** (some say X, others Y) = Investigate
- **Surprises** (didn't expect this) = Important insight
- **Specifics** (concrete suggestions) = Actionable

Select 2-3 quotes per theme:

- Brief (2-3 sentences max)
- Anonymise (remove names)
- Illustrate the point clearly

STEP 4: COMPARE DIFFERENT GROUPS

Theme	Participants	Staff	Partners	Match?
Waiting times	"Too long"	"We're at capacity"	"Clients frustrated"	✓ All agree = Real issue
Outcomes	"Helped me a lot"	"Not seeing employment results"	"Unclear what clients gain"	✗ Disagree = Investigate

When views diverge, investigate why.

STEP 5: INTERPRET CRITICALLY

Ask yourself:

Who responded? <50% response = Bias risk. Who's missing?

Being polite? If zero negative comments, be suspicious (especially if staff collected data)

External factors? Pandemic? Funding cuts? Policy change affecting satisfaction?

Small sample? n<30 = Differences might be chance not pattern

Don't: Cherry-pick positives. Dismiss criticism. Over-interpret n=3 comments.

Do: Look for patterns across multiple sources. Be honest about limitations.

STEP 6: WRITE YOUR FINDINGS (1-2 PAGES)

Strengths (2-3 bullet points): "Staff support highly valued (mean 4.6/5, 91% positive). Participants described staff as *'respectful, patient, really listened'*."

Areas for improvement (2-4 bullet points): "Access barriers mentioned by 15% (n=7). Specifically transport difficulties and inconvenient timing: *'Hard to get there on public transport'*."

Surprises: Anything unexpected (positive or negative)

Divergent views: Where stakeholders disagree

Priority actions: 3-5 concrete next steps

STEP 7: DECIDE WHAT TO DO

Quick wins (do now): Example: Clarify programme information (confusion mentioned in feedback) → Action within 1-3 months

Medium-term (plan it): Example: Address transport barriers (change times or locations) → Action within 6-12 months

Strategic (longer-term): Example: Redesign programme model → Action within 12-24 months; discuss with funder

Cannot action: Example: Cannot change eligibility (funder requirement) → Be transparent. Explain what you CAN do instead.

STEP 8: CLOSE THE LOOP

Use "You Said, We Did" template (B.3.4) to tell stakeholders:

- What you heard (main themes)
- What you're doing about it
- What you can't change and why
- When they'll be asked again

Share with: Staff, participants, partners, funders

B.4: COST-EFFECTIVENESS ANALYSIS

Purpose: This annex provides integrated tools for programmes implementing Cost-Effectiveness Analysis (CEA) comparing alternative delivery approaches. CEA compares costs and outcomes of different interventions identifying which achieves outcomes most efficiently.

Who uses this: Medium-scale programmes with evaluation capacity, or commissioners specifying requirements for external evaluators. For complete methodological guidance, see Section 3.3.1.

B.4.1 CEA Planning and Scope Definition

Programme: _____

Evaluation lead: _____

Date: _____

Scope Definition

Element	Specification	Justification
Perspective	<input type="checkbox"/> Programme <input type="checkbox"/> Funder <input type="checkbox"/> Public sector <input type="checkbox"/> Societal	
Time horizon	_____ years	
Discount rate	_____ % (3% EU standard, 3.5% UK standard)	
Target population		
Primary outcome measure		
Alternatives compared	Alternative A: Alternative B: Alternative C (if applicable):	
Comparator rationale	Why these alternatives?	

Data Availability Assessment

Data Required	Available?	Source	Quality
Cost data (Alternative A)	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> Good <input type="checkbox"/> Adequate <input type="checkbox"/> Poor
Cost data (Alternative B)	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> Good <input type="checkbox"/> Adequate <input type="checkbox"/> Poor
Outcome data (Alternative A)	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> Good <input type="checkbox"/> Adequate <input type="checkbox"/> Poor
Outcome data (Alternative B)	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> Good <input type="checkbox"/> Adequate <input type="checkbox"/> Poor
Participant numbers	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> Good <input type="checkbox"/> Adequate <input type="checkbox"/> Poor



B.4.2 Integrated CEA Calculation Tool

Instructions: Complete this integrated spreadsheet calculating costs, outcomes, cost-effectiveness ratios, and sensitivity analysis. Use one spreadsheet per CEA with multiple sheets as indicated below.

Template 12 – CEA Cost Data

Cost Category	Alternative A	Alternative B	Alternative C	Notes/Sources
Programme Costs				
Staff salaries (direct delivery)				FTE × salary + 25% on-costs
Staff salaries (management/admin)				FTE × salary + 25% on-costs
Training and supervision				Hours × hourly rate
Facilities and rent				Sq metres × rate, proportionally allocated
Materials and consumables				Per participant × participants
Equipment and IT				Annualized if capital cost
Travel and transport				Actual or standard mileage rate
Total Programme Costs				
Participant Costs (if societal perspective)				
Participant time				Hours × opportunity cost
Participant transport				Actual or estimated
Total Participant Costs				
TOTAL COSTS (undiscounted)				
Discount factor (if multi-year)				$(1/(1+r)^t)$
TOTAL COSTS (discounted)				
Number of participants				
Cost per participant				

Template 13 – CEA Outcome Data

Outcome Measure	Alternative A	Alternative B	Alternative C	Notes/Sources
Primary outcome				
Total outcomes achieved				Count or sum
Number of participants				
Outcomes per participant				
Data quality indicators				
Baseline completion rate	___%	___%	___%	Should be >80%
Follow-up completion rate	___%	___%	___%	Should be >70%
Attrition rate	___%	___%	___%	Should be similar across alternatives

Template 14 - Cost-Effectiveness Ratios

Metric	Alternative A	Alternative B	Alternative C	Interpretation
Total cost (discounted)	[from Sheet 1]	[from Sheet 1]	[from Sheet 1]	
Total outcomes	[from Sheet 2]	[from Sheet 2]	[from Sheet 2]	
CER (Cost per outcome)				Lower is more cost-effective
Incremental analysis (compared to Alternative A)				
Incremental cost (ΔC)	Reference (0)			
Incremental outcome (ΔE)	Reference (0)			

ICER	Reference			Cost per additional outcome unit
Dominated?	No	<input type="checkbox"/> Yes (more costly, less effective)	<input type="checkbox"/> Yes	
Dominant?		<input type="checkbox"/> Yes (less costly, more effective)	<input type="checkbox"/> Yes	

ICER Interpretation:

- Negative ICER = Alternative less costly AND more effective (DOMINANT - clearly prefer)
- Positive ICER = Alternative more costly AND more effective (decision depends on willingness to pay for additional outcomes)
- Dominated alternative = More costly AND less effective (reject)

SHEET 4: Sensitivity Analysis

Test how conclusions change when varying key assumptions:

Template 15 – CEA Sensitivity Analysis

Parameter Varied	Base Case	Low Estimate	High Estimate	CER Range Alternative A	CER Range Alternative B	Conclusion Robust?
Staff costs		-20%	20%			<input type="checkbox"/> Yes <input type="checkbox"/> No
Outcome rate		-10%	10%			<input type="checkbox"/> Yes <input type="checkbox"/> No
Overhead allocation	Method used:	Alternative method 1	Alternative method 2			<input type="checkbox"/> Yes <input type="checkbox"/> No
Time horizon	___ years	___ years	___ years			<input type="checkbox"/> Yes <input type="checkbox"/> No
Discount rate	___%	0% (no discounting)	5%			<input type="checkbox"/> Yes <input type="checkbox"/> No

Sensitivity Analysis Summary:

Findings are ROBUST (preferred alternative remains preferred across plausible variations) or SENSITIVE (conclusion changes with parameter variation—flag which parameters matter most for decision-makers).

B.4.3 Quality Assurance Checklist

Use this checklist before finalizing CEA to catch common errors.

Programme: _____

Evaluator: _____

Date: _____

Scope Definition

- Perspective clearly stated (programme/funder/public sector/societal)
- Perspective consistently applied (all relevant costs and outcomes from that perspective counted)
- Time horizon justified (long enough for costs and outcomes to emerge)
- Discount rate applied correctly if multi-year (3% EU or 3.5% UK standard)
- Comparator appropriate (realistic alternative, not just "do nothing")
- Alternatives genuinely comparable (similar target population, context, delivery period)

Cost Identification

- All staff costs captured (salaries + 20-30% on-costs)
- Direct activity costs included (materials, participant costs, venue)
- Overhead costs allocated proportionally (management, office space, IT, utilities)
- Capital costs annualised if applicable (equipment, vehicles over useful life)
- No double-counting (each cost counted once only)
- Donated resources valued at market rate (volunteers, pro-bono space, in-kind)
- Cost allocation method documented and consistent across alternatives
- Actual costs used where available (not just budgeted amounts)

Outcome Measurement

- Primary outcome clearly defined (specific, measurable)
- Outcomes measured identically across all alternatives (same definition, tools, timing)
- Valid measurement tools used (validated scales where available)
- Follow-up rates adequate (>70% for credible outcome data)
- Attrition documented and similar across alternatives
- Outcome timing appropriate (measured when meaningful change expected)

Calculations

- CER calculated correctly (Total cost ÷ Total outcomes)
- ICER calculated correctly (Incremental cost ÷ Incremental outcomes)
- Discounting applied correctly if multi-year (both costs and outcomes discounted)

- Spreadsheet formulas checked (spot-check calculations manually)
- Cost per participant calculated for context

Sensitivity Analysis

- Sensitivity analysis conducted (varied key parameters)
- Key parameters tested: Staff costs ($\pm 20\%$), Outcome rates ($\pm 10\%$), Overhead allocation (alternative methods), Time horizon (shorter/longer), Discount rate (0%, 3%, 3.5%, 5%)
- Results documented (which parameters change conclusion?)
- Robustness assessed (do findings hold across plausible parameter ranges?)

Presentation

- Results tables complete: Costs by alternative, Outcomes by alternative, Cost-effectiveness ratios
- Key finding stated clearly in plain language (which alternative most cost-effective, by how much?)
- Sensitivity analysis results described (robust or sensitive to assumptions?)
- Limitations explicitly stated: Data limitations, Exclusions, Uncertainty, Threats to validity
- Accessible for decision-makers (avoid jargon, explain implications)
- Assumptions documented (overhead allocation method, valuation approach, discount rate)

Common Errors Check

- NOT comparing incomparable programmes (alternatives serve similar populations with similar aims)
- NOT mixing perspectives (e.g., counting only programme costs but all societal outcomes)
- NOT using too-short time horizon (that misses delayed costs or outcomes)
- NOT presenting single point estimate without sensitivity analysis
- NOT claiming causation without appropriate comparison design (CEA shows association not causation unless comparison group controls for selection)
- NOT ignoring context (cost-effectiveness in one setting may not transfer to another)

B.5: MULTI-CRITERIA DECISION ANALYSIS

Purpose: This annex provides integrated tools for programmes implementing Multi-Criteria Decision Analysis (MCDA) evaluating options against multiple criteria when no single metric captures all relevant considerations. MCDA makes explicit the values and trade-offs underlying programme decisions.

Who uses this: Programmes facing complex decisions with multiple competing objectives, or commissioners specifying MCDA requirements. For complete methodological guidance, see Section 3.3.2.

B.5.1 MCDA Planning and Criteria Development

Decision problem: _____

Facilitator/Analyst: _____

Date: _____

Decision Context

What decision needs to be made:

Options being evaluated:

- Option A: _____
- Option B: _____
- Option C: _____
- Option D (if applicable): _____

Stakeholders to involve:

- Programme staff Management Service users/representatives
- Funders Partners Governance board Other: _____

Timeline:

- Criteria identification workshop: _____
- Weighting workshop: _____
- Performance scoring: _____
- Analysis and presentation: _____

Criteria Identification and Refinement

Instructions: Brainstorm comprehensive criteria, then refine to 5-12 distinct measurable criteria.

Brainstormed Criterion	Category	Measurable?	Distinct?	Decision
	<input type="checkbox"/> Outcomes <input type="checkbox"/> Cost <input type="checkbox"/> Feasibility <input type="checkbox"/> Equity <input type="checkbox"/> Other	<input type="checkbox"/> Yes <input type="checkbox"/> No	<input type="checkbox"/> Yes <input type="checkbox"/> No	<input type="checkbox"/> Keep <input type="checkbox"/> Refine <input type="checkbox"/> Remove

Final Criteria (5-12 maximum):

#	Criterion Name	Definition	How Measured	Scale
1				1-5 where 1=poor, 5=excellent
2				
3				
4				
5				
6				

Quality checks:

- All criteria distinct (can score high on one, low on another)
- All criteria measurable for options being evaluated
- No double-counting (e.g., "participant satisfaction" and "service quality" may measure same thing)
- No criteria universally satisfied by all options (no discriminating power)
- Number manageable (5-12 criteria)

B.5.2 Integrated MCDA Calculation Tool

Instructions: Complete this integrated tool weighting criteria, scoring options, calculating weighted scores, and conducting sensitivity analysis.

SHEET 1: Weights and Performance Matrix

Weighting Method Used: Direct Allocation (distribute 100 points) Swing Weighting (rank importance of moving from worst to best)

Template 16 - MCDA Weights and Performance Matrix

Criterion	Weight (must sum to 1.0)	Option A Score (1-5)	Option B Score (1-5)	Option C Score (1-5)	Evidence/ Rationale
1. _____					
2. _____					
3. _____					
4. _____					
5. _____					
6. _____					
TOTAL	1.00				
WEIGHTED SCORE					
RANK					

SHEET 2: Sensitivity Analysis - Weight Variations

Test how conclusions change when varying weights:

Template 17 – MCDA Sensitivity Analysis A

Scenario	Weight Adjustments	Option A Score	Option B Score	Option C Score	Preferred Option	Robust?
Base case	[Weights from Sheet 1]					
Emphasize outcomes	Outcome criteria +10%, others adjusted proportionally					<input type="checkbox"/> Yes <input type="checkbox"/> No



Emphasize cost	Cost criterion +10%, others adjusted					<input type="checkbox"/> Yes <input type="checkbox"/> No
Emphasize equity	Equity criterion +10%, others adjusted					<input type="checkbox"/> Yes <input type="checkbox"/> No
Stakeholder Group X weights	[If stakeholders weighted differently]					<input type="checkbox"/> Yes <input type="checkbox"/> No
Equal weights (all criteria weighted equally)	All weights = 1/n					<input type="checkbox"/> Yes <input type="checkbox"/> No

Sensitivity Summary: Preferred option REMAINS SAME CHANGES across scenarios

SHEET 3: Sensitivity Analysis - Score Variations

Test how conclusions change when varying performance scores:

Template 18 – MCDA Sensitivity Analysis B

Scenario	Score Adjustments	Option A Score	Option B Score	Option C Score	Preferred Option	Robust?
Base case	[Scores from Sheet 1]					
Optimistic for A	A scores +1 where uncertain					<input type="checkbox"/> Yes <input type="checkbox"/> No
Pessimistic for A	A scores -1 where uncertain					<input type="checkbox"/> Yes <input type="checkbox"/> No
Optimistic for B	B scores +1 where uncertain					<input type="checkbox"/> Yes <input type="checkbox"/> No
Pessimistic for B	B scores -1 where uncertain					<input type="checkbox"/> Yes <input type="checkbox"/> No

Robustness Assessment:

- Findings are ROBUST (preferred option remains preferred across plausible variations)
- Findings are SENSITIVE (preferred option changes with variations—flag which parameters critical)

B.5.3 Quality Assurance Checklist

Use this checklist before finalising MCDA to ensure quality.

Decision: _____

Analyst: _____

Date: _____

Criteria Quality

- Criteria comprehensive (all important decision dimensions captured: outcomes, costs, feasibility, equity, sustainability)
- Criteria distinct (minimal overlap—can score high on one criterion but low on another)
- Criteria measurable (can actually assess how options perform on each criterion)
- Number manageable (5-12 criteria; fewer than 5 too limited, more than 12 unwieldy)
- No double-counting (highly correlated criteria avoided—e.g., "participant satisfaction" and "service quality" may measure same thing)
- No unmeasurable criteria (all criteria can be scored for options being considered)
- No universally satisfied criteria (if all options perform identically, criterion doesn't discriminate)

Weighting Process

- Stakeholder participation documented (who participated in weighting?)
- Diverse perspectives included (staff, service users, management, funders as appropriate)
- Weighting method documented (Direct Allocation or Swing Weighting?)
- Individual weights collected before group discussion (prevents groupthink)
- Weights sum to 1.0 (or 100%)
- No single criterion dominates (typically no criterion >35-40% unless genuinely overriding)
- Weight distribution reflects genuine priorities (not artificially equal)
- Areas of disagreement documented (value conflicts made transparent)
- Rationale for weights documented (why these weights reflect stakeholder values?)

Performance Scoring

- Scoring method documented (quantitative data, expert judgment, or mixed)
- Evidence basis clear for each score (data sources or judgment rationale documented)
- Scoring scales consistent (same scale applied to all criteria and options)
- Scoring process participatory (not analyst alone—stakeholders involved)

- Areas of uncertainty documented (where confidence in scores low)
- No bias toward predetermined option (scoring based on evidence not preferences)

Calculation and Analysis

- Weighted scores calculated correctly ($=\sum (\text{Weight} \times \text{Score})$ for each option)
- Spreadsheet formulas verified (spot-check calculations manually)
- Rankings clear (which option scores highest)
- Margin of superiority calculated (how much better is preferred option?)

Sensitivity Analysis

- Weight sensitivity tested (varied weights to see if conclusions robust)
- Score sensitivity tested (varied uncertain scores within plausible ranges)
- Results documented (which parameters matter most?)
- Robustness assessed (does preferred option remain preferred across plausible variations?)
- Critical factors identified (where conclusion sensitive, flagged for decision-makers)

Stakeholder Engagement

- Participation genuine not tokenistic (stakeholders shaped process, not just consulted)
- Diverse perspectives included (not just management or analyst view)
- Process transparent (stakeholders understand how analysis conducted)
- Disagreements acknowledged and documented (not forced into false consensus)
- Results shared with participants before finalising
- Feedback incorporated where appropriate
- Decision-makers involved at key stages (criteria selection, weighting, interpreting results)

Presentation and Documentation

- Decision problem clearly stated
- Options described adequately
- Criteria and weights presented in table with rationale
- Performance scores shown with evidence basis
- Visual presentation included (bar chart of weighted scores, radar chart showing profiles)
- Interpretation provided: Preferred option, margin, strengths/weaknesses, trade-offs
- Sensitivity analysis results described (robust or sensitive?)
- Limitations acknowledged (data quality, judgment uncertainty, scope)
- Recommendation contextualised (MCDA results PLUS broader considerations)

Common Errors Check

- NOT criteria overlap (multiple criteria measuring same underlying dimension)
- NOT predetermined conclusions (manipulating criteria or weights to justify decision already made)
- NOT spurious precision (presenting scores to many decimal places suggesting false precision)
- NOT dominating criteria (one criterion weighted so heavily it determines result regardless of others)
- NOT unmeasurable criteria (criteria sound important but cannot actually be assessed)
- NOT inadequate stakeholder engagement (analyst-led with token consultation)
- NOT mechanical application (MCDA results applied without considering implementation realities, relationships, political context)

B.6: Social Return on Investment (SROI) Commissioning Specification

B.6.1 SROI Requirements and Resource Implications

What SROI requires:

Methodological requirements:

- Stakeholder engagement throughout evaluation process (outcome identification, proxy validation, interpretation)
- Theory of Change explicitly linking activities to outcomes for each stakeholder group
- Financial proxy valuation with stakeholder validation
- Mandatory impact adjustments: deadweight (what would happen anyway), attribution (contribution of other factors), displacement (benefits to some at expense of others), drop-off (outcome decay over time)—all four applied to every outcome
- Sensitivity analysis presenting SROI as range not single figure
- Compliance with Social Value International Seven Principles

Data requirements:

- Quantified outcomes for multiple stakeholder groups (participants, families, staff, partners, public services)
- Evidence supporting impact adjustment estimates (comparison data, academic literature, stakeholder knowledge)
- Cost records comprehensive enough for calculating total investment

Expertise requirements:

- SROI-accredited practitioner (Social Value International accreditation or equivalent)
- Facilitation skills enabling genuine stakeholder participation (not tokenistic consultation)
- Financial proxy valuation expertise
- Impact adjustment methodology understanding

Resource implications:

- Specialist time: Typically 2-4 months depending on stakeholder engagement complexity
- External SROI practitioner fees: Vary by country and practitioner experience—obtain estimates from accredited practitioners
- Stakeholder engagement costs: Workshop facilitation, participant interviews, validation sessions
- Timeline: 3-6 months total including stakeholder engagement, data collection, analysis, reporting

Contexts using SROI:

- Social investment requiring SROI evidence (impact bonds, venture philanthropy)
- Funders explicitly requesting SROI alongside outcome monitoring
- Programmes where stakeholder voice central to values and multiple stakeholder groups experience outcomes
- Contexts where conventional economic valuation inadequate but stakeholders can articulate value

Validity threats when requirements not met:

- Tokenistic stakeholder engagement → analyst-imposed outcomes and valuations lacking legitimacy
- Incomplete impact adjustments → inflated SROI ratios over-claiming programme contribution
- No sensitivity analysis → single SROI figure suggesting false precision
- Lack of accredited practitioner → non-compliance with Social Value International standards, inappropriate proxy selection

B.6.2 Scope Definition Requirements

Commissioners must specify:

Specification	What to Define
SROI type	Evaluative SROI (retrospective—actual impact achieved) or Forecast SROI (prospective—predicted impact)
Stakeholder groups	Which groups to engage? Participants, families, staff, partners, public services, employers, community
Materiality threshold	How to prioritize outcomes? (Include outcomes affecting >10% stakeholders OR valued highly by any group)
Impact adjustments	Require all four mandatory adjustments: deadweight, attribution, displacement, drop-off
Time horizon	Typical 3-5 years for social services, justify based on outcome duration
Available data	Outcome data, cost records, stakeholder access for interviews/workshops
Budget and timeline	Evaluation budget, timeline (SROI typically 2-4 months), reporting deadline

B.6.3 Quality Standards Checklist

Verify the following:

Stakeholder Engagement:

- Genuine participatory process (not tokenistic consultation)
- Multiple stakeholder groups consulted (participants, families, staff minimum)
- Stakeholders identified material outcomes (not analyst-imposed)
- Stakeholders validated financial proxy selections
- Engagement documented (who participated, when, how input influenced analysis)

Theory of Change:

- Explicit Theory of Change links activities to outcomes for each stakeholder group
- Outcomes identified through stakeholder process (materiality principle applied)

Financial Proxies:

- Proxies justified and stakeholder-validated
- Proxy sources cited (Global Value Exchange, Social Value UK, academic studies)
- Proxies appropriate for context (UK proxies not directly applied in other EU countries without adjustment)

Impact Adjustments:

- Deadweight** calculated (what would have happened anyway) with evidence/rationale
- Attribution** calculated (contribution of other factors) with evidence/rationale
- Displacement** assessed (programme benefits some at expense of others)
- Drop-off** calculated (outcome decay over time) with evidence/rationale
- All four adjustments applied to every outcome

Sensitivity Analysis:

- SROI presented as range not single figure (optimistic and conservative scenarios)
- Key assumptions varied systematically
- Robustness assessed

Transparency and Compliance:

- Methods enable verification by independent analyst
- Compliance with Social Value International Principles verified
- Limitations acknowledged
- Independent assurance obtained if high-stakes (e.g., social impact bond)

B.6.4 Reference Standards

Social Value International Principles: All SROI must comply with seven principles: (1) Involve stakeholders, (2) Understand what changes, (3) Value the things that matter, (4) Only include what is material, (5) Do not over-claim, (6) Be transparent, (7) Verify the result.

Financial Proxy Sources:

- Global Value Exchange (primary resource for EU contexts, international proxies)
- Social Value UK Portal (UK-specific proxies)
- HACT Social Value Bank (housing and wellbeing proxies UK)
- National wellbeing surveys (country-specific subjective wellbeing data)

Impact Adjustment Evidence:

- Comparison group data (strongest evidence for deadweight and attribution)
- Academic literature on natural improvement rates and programme effectiveness
- Stakeholder consultation (displacement, attribution to partners)
- Prior evaluations of similar programmes

B.7: Cost-Benefit Analysis Commissioning Specification

Purpose: This annex provides commissioning specifications for Cost-Benefit Analysis (CBA). CBA require specialist expertise and cannot be implemented reliably by programme staff alone. Organisations should commission external specialists using these specifications to define scope, ensure quality, and verify compliance with EU/UK standards.

Who uses this: Commissioners defining requirements for external evaluators, or quality assurance reviewers verifying specialist work. For methodological principles, see Section 3.4.1 (CBA).

B.7.1 CBA Requirements and Resource Implications

What CBA requires:

Methodological requirements:

- Comprehensive benefit identification across all outcome domains (health, employment, education, criminal justice, welfare, wellbeing)
- Monetary valuation of identified benefits using established economic methods
- Comprehensive cost accounting from specified perspective (societal, public sector, or funder)
- Discounting of future costs and benefits to present value (3% EU standard, 3.5% UK standard)
- Sensitivity analysis testing robustness to assumptions
- Distributional analysis identifying who gains and who loses

Data requirements:

- Quantified outcomes with sufficient reliability for benefit estimation
- Cost records covering all relevant cost categories
- Administrative data access where benefits depend on health service use, criminal justice contact, or welfare receipt
- Follow-up data capturing when benefits materialize (may extend multiple years post-intervention)

Expertise requirements:

- Specialist with demonstrated CBA expertise who understands monetary valuation methods, discounting, distributional analysis, and sensitivity testing

Resource implications:

- Specialist time: Typically 3-6 months depending on complexity and data availability
- External specialist fees: Vary substantially by country and specialist experience— obtain estimates from multiple qualified specialists
- Data collection costs: If new data collection required (participant surveys, administrative data linkage)
- Timeline: 6-12 months total from commissioning to final report including data collection, analysis, stakeholder review

B.7.2 Scope Definition Requirements

Commissioners must specify the following to evaluators:

Specification	What to Define	Example
Perspective	Whose costs and benefits counted?	Societal (most comprehensive), Public Sector (government costs/benefits only), Funder (funder's costs/benefits)
Time horizon	Analysis period capturing costs and benefits	1-2 years (short-term interventions), 3-5 years (medium-term), 10+ years (life-changing interventions). Justify based on when benefits materialize.
Comparator	What is programme compared against?	Business-as-usual (current practice), Do-minimum (minimal intervention), Alternative delivery model
Primary benefit categories	What outcomes to monetise?	Employment outcomes, health service use, criminal justice contact, education progression, welfare dependency, wellbeing improvements
Available data	What data exists or can be collected?	Cost records (specify completeness), Outcome data (specify measures, sample size, timing), Administrative data access (health, justice, welfare records), Comparison group availability
Budget and timeline	Resources for evaluation	Evaluation budget: €_____, Timeline: Start _____ Complete _____, Reporting deadline: _____
Deliverables	What outputs required?	Technical report, Executive summary (2-3 pages), Presentation to stakeholders, Sensitivity analysis showing robustness

Additional specifications:

- Will evaluators have direct access to programme participants for surveys/interviews?
- Are Data Sharing Agreements in place for administrative data linkage?
- Are GDPR compliance and ethics approval arranged, or must evaluator obtain?
- What non-monetised outcomes should be reported qualitatively alongside BCR?

B.7.3 EU/UK Valuation Standards and Reference Values

Purpose: This section provides commissioners with knowledge of what CAN be monetised, what valuation approaches exist, and WHERE to find current values. This enables commissioners to specify comprehensive benefit identification and verify evaluators used appropriate sources.

CRITICAL: Do NOT reproduce actual monetary values here—they change annually. Instead, reference authoritative sources providing current values.

Standard Discount Rates

- **EU Standard:** 3% social discount rate (EU Better Regulation Guidelines, Annex III)
- **UK Standard:** 3.5% discount rate (HM Treasury Green Book, Annex 6)
- **Application:** All future costs and benefits discounted to present value
- **Long-term discounting:** For impacts beyond 30 years, declining discount schedule may apply (consult EU Better Regulation Vademecum or UK Green Book Annex 6)

Common Benefit Categories with Established Valuation Approaches

1. Employment Outcomes

- **What can be valued:** Earnings gains, income tax revenue, National Insurance contributions, welfare benefit savings (unemployment benefits, housing benefits, disability benefits)
- **Valuation approach:** Market prices (actual wages), fiscal impacts (tax/benefit schedules)
- **Current values source:**
 - Eurostat Labour Force Survey (wage data by country, sector, skill level)
 - National statistical offices (earnings distributions)
 - OECD Employment Outlook (international comparisons)
 - National tax authorities (tax rates, benefit schedules)
- **Key consideration:** Distinguish gross earnings (participant benefit) from fiscal impact (government benefit) to avoid double-counting

2. Health Service Use

- **What can be valued:** GP visits, hospital admissions, emergency department attendances, outpatient appointments, mental health services, substance abuse treatment, preventive care
- **Valuation approach:** Unit costs (average cost per service episode)
- **Current values source:**
 - National health service unit cost databases (e.g., NHS Reference Costs UK, DRG tariffs other EU countries)
 - WHO-CHOICE (cost-effectiveness thresholds by country)
 - National health ministries publish annual reference cost schedules
- **Key consideration:** Use national unit costs from country where services delivered—do not transfer values across borders without adjustment

3. Quality-Adjusted Life Years (QALYs)

- **What can be valued:** Health-related quality of life improvements, life expectancy gains
- **Valuation approach:** Willingness-to-pay for QALY gains
- **Current values source:**
 - UK: £70,000 per QALY (2020/21 prices, HM Treasury Green Book Annex 1, updated for inflation annually)
 - Netherlands: €80,000 per QALY
 - Other EU countries: Consult national health technology assessment bodies
 - WHO: Uses GDP per capita thresholds (highly cost-effective if $<1 \times$ GDP per capita per DALY averted)
- **Key consideration:** QALY valuation contentious—some stakeholders object to monetising life/health. Consider presenting QALYs gained alongside BCR without forced monetisation.

4. Criminal Justice Contact

- **What can be valued:** Police contact, court proceedings, incarceration, probation, victim costs, re-offending prevention
- **Valuation approach:** Unit costs (average cost per criminal justice event)
- **Current values source:**
 - UK: Home Office Economic and Social Costs of Crime (updated periodically)
 - EU: National ministries of justice publish criminal justice cost data
 - Academic literature: systematic reviews of criminal justice costs
- **Key consideration:** Distinguish costs to justice system (police, courts, prisons) from costs to victims and society—comprehensive CBA includes all

5. Education Outcomes

- **What can be valued:** Educational attainment improvements, lifetime earnings impact, productivity gains
- **Valuation approach:** Human capital approach (present value of lifetime earnings differences by qualification level)
- **Current values source:**
 - OECD Education at a Glance (lifetime earnings by education level)
 - National labour force surveys (earnings by qualification)
 - Academic literature on education returns (Mincer equations estimating wage premium per year of education)
- **Key consideration:** Education benefits accrue over decades—requires long time horizon and careful discounting

6. Wellbeing Improvements

- **What can be valued:** Life satisfaction changes, subjective wellbeing improvements
- **Valuation approach:** Compensating income equivalents (how much income change produces equivalent wellbeing change)
- **Current values source:**
 - UK: HM Treasury Wellbeing Guidance (supplementary Green Book guidance)
 - Academic literature: Fujiwara & Campbell (2011) "Valuation Techniques for Social Cost-Benefit Analysis"
 - Layard et al. research on wellbeing valuation
- **Key consideration:** Wellbeing valuation highly contested. Many evaluators prefer reporting wellbeing improvements qualitatively alongside fiscal BCR rather than monetising subjective wellbeing.

7. Social Care Services

- **What can be valued:** Domiciliary care, residential care, respite care, day services, assistive technology, informal carer time
- **Valuation approach:** Unit costs (average cost per care hour/week)
- **Current values source:**
 - UK: Personal Social Services Research Unit (PSSRU) Unit Costs of Health and Social Care (annual publication)
 - EU countries: National social care cost schedules from ministries
 - Informal care: Opportunity cost method (carer's foregone wages) or replacement cost method (market rate for equivalent care)
- **Key consideration:** Social care costs vary dramatically across countries—use country-specific sources

Valuation Approaches Overview

Market Prices: Where competitive markets exist, market prices reflect social value. Use for: employment (wages), goods/services purchased, capital assets.

Revealed Preference: Infer values from observed behaviour. Use for: travel time (wage rates), housing (hedonic pricing), environmental amenities (property values).

Stated Preference: Ask people directly what they value. Methods: Contingent Valuation (willingness to pay/accept), Discrete Choice Experiments (trade-offs between attributes). Use cautiously—stated often exceeds actual willingness to pay.

Transfer Method: Use established values from prior studies. Most pragmatic approach for social services CBA. Requirements: Similar context, adjustment for inflation, adjustment for purchasing power across countries if transferring internationally.

Human Capital Approach: Value individuals' productivity contributions. Use for: education benefits (lifetime earnings), mortality (value of statistical life based on productivity loss). Criticized as undervaluing non-working populations.

Compensating Income Equivalent: How much income change produces equivalent wellbeing change. Use for: wellbeing impacts, subjective outcomes. Derived from large-scale surveys relating income and wellbeing.

What NOT to Monetise

Some outcomes resist credible monetary valuation or stakeholders fundamentally object to monetisation:

- Human dignity, autonomy, voice as primary programme objectives (not means to other ends)
- Justice, rights, fairness where programme aims for equity not efficiency
- Cultural preservation, community cohesion, social capital where no established valuation methods
- Outcomes where monetisation would undermine stakeholder support or programme legitimacy

Recommended approach: Present these outcomes qualitatively alongside BCR. Example: "BCR 3.1 based on fiscal benefits (health, justice, welfare savings). Additionally, 78% participants reported restored dignity and 65% increased community engagement."

Sources Summary

EU Standards:

- EU Better Regulation Guidelines, Annex III: CBA Vademecum (<https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation>)
- European Commission Better Regulation Toolbox, Tool #61: Cost-Benefit Analysis

UK Standards:

- HM Treasury Green Book: Central Government Guidance on Appraisal and Evaluation (<https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government>)
- HM Treasury Supplementary Guidance on Wellbeing Valuation

National Sources:

- National health ministries: Unit cost schedules, QALY thresholds
- National justice ministries: Criminal justice cost data
- National statistical offices: Labour force surveys, earnings data
- Health technology assessment bodies: QALY valuation, cost-effectiveness thresholds

Academic Resources:

- Boardman et al. (2017) "Cost-Benefit Analysis: Concepts and Practice" (comprehensive textbook)
- Fujiwara & Campbell (2011) "Valuation Techniques for Social Cost-Benefit Analysis" (UK Treasury guide, free online)
- Journal of Benefit-Cost Analysis, Applied Health Economics and Health Policy (peer-reviewed research)

B.7.4 Optimism Bias and Uncertainty Standards

Optimism bias: Programmes systematically overestimate benefits and underestimate costs. This is human nature, not dishonesty. Evaluators must adjust for this.

EU approach: Mandatory sensitivity analysis with pessimistic scenarios testing whether conclusions robust to optimistic assumptions being wrong.

UK approach: Explicit optimism bias adjustments applied upfront. UK Treasury Green Book Annex 5 provides tables showing typical optimism bias by intervention type:

- Non-standard civil buildings: 24% cost increase
- Standard IT projects: 41% cost increase, 41% benefit shortfall
- Non-standard equipment: 54% cost increase
- Refer to UK Green Book Annex 5, Table 1 for complete adjustments by project type

Commissioner requirement: Evaluators must either:

1. Apply standard optimism bias adjustments from UK Green Book (if UK context) or demonstrate adjustments based on comparable programmes (EU context), OR
2. Conduct comprehensive sensitivity analysis showing BCR remains >1.0 under pessimistic assumptions (benefits 20-30% lower than base case, costs 20-30% higher than base case)

Uncertainty management: CBA involves numerous uncertain parameters (benefit magnitudes, unit costs, discount rate, time horizon). Commissioners should require:

- **One-way sensitivity analysis:** Vary each key parameter individually showing effect on BCR/NPV
- **Scenario analysis:** Optimistic, expected, pessimistic scenarios with BCR/NPV for each
- **Threshold analysis:** Identify "switching values" where BCR crosses 1.0 (e.g., "If employment rate falls below 45%, $BCR < 1.0$. Observed: 63%.")
- **Probabilistic analysis (if feasible):** Monte Carlo simulation varying multiple parameters simultaneously showing probability $BCR > 1.0$
- **Reporting requirement:** Present BCR as range not point estimate. Example: "BCR 2.8 under base assumptions, ranging from 1.9 (pessimistic) to 4.2 (optimistic)."

B.7.5 Quality Standards Checklist

Commissioners should verify the following standards:

Scope and Perspective: Societal perspective applied consistently (or narrower perspective justified explicitly)

- All relevant costs and benefits from stated perspective counted (not cherry-picked)
- Time horizon justified based on when benefits materialize
- Comparator appropriate (realistic alternative, not just "do nothing")

Benefit Identification and Valuation: Comprehensive benefit identification using Theory of Change as framework

- All significant benefit categories addressed (even if not monetised)
- Monetary valuation methods justified for each benefit category
- Valuation sources cited (unit cost databases, academic literature, prior studies)
- Transfer values adjusted for inflation and context
- Non-monetised outcomes reported explicitly alongside BCR (not hidden)

Cost Accounting: All cost categories identified systematically

- Programme delivery costs (staff, materials, overheads, capital) included
- Participant costs included if societal perspective (time, transport)
- System costs included where relevant (public services, partners)
- Cost allocation methods documented
- No double-counting (each cost counted once only)

Discounting: 3% (EU) or 3.5% (UK) discount rate applied (or alternative justified)

- Discounting applied consistently to both costs and benefits
- Long-term impacts (>30 years) use declining discount schedule if relevant

Impact Adjustments: Deadweight addressed (what would have happened anyway without programme)

- Attribution addressed (contribution of other factors alongside programme)
- Displacement addressed if relevant (programme benefits some at expense of others)
- Adjustments evidence-based (comparison data, academic literature) not arbitrary

Uncertainty Management: Optimism bias addressed (explicit adjustments OR comprehensive sensitivity analysis)

- Sensitivity analysis conducted varying key parameters
- Scenario analysis (optimistic, expected, pessimistic) presented
- Threshold analysis identifies switching values
- Results presented as range not single point estimate

Distributional Analysis: Who gains and who loses identified

- Benefits disaggregated by participant characteristics where data permit
- Equity implications discussed (progressive, neutral, or regressive distribution)

Transparency and Documentation: Methods documented enabling replication by independent analyst

- All assumptions stated explicitly with justification
- Data sources specified for all costs and benefits
- Limitations acknowledged (data gaps, valuation uncertainties, threats to causal inference)
- Sufficient detail for peer review

Independence and Credibility: Evaluator credentials verified (Qualifications in economics or public policy, published CBA work, relevant experience)

- Independent quality assurance obtained for high-stakes CBA (second health economist reviews valuation choices, analytical approach)
- No conflicts of interest (evaluator not financially invested in programme success)

B.8: Advanced Causal and Complexity Evaluation Commissioning Specifications

Purpose: This annex provides commissioning specifications for Quasi-Experimental Design (QED), Randomised Controlled Trials (RCT), and Realist Evaluation. These methods require specialist expertise—causal inference specialists, trial methodologists, skilled qualitative researchers—and cannot be implemented reliably by programme staff alone. Organisations should commission external specialists using these specifications to define scope, ensure quality, and verify compliance with methodological standards.

Who uses this: Commissioners defining requirements for external evaluators, or quality assurance reviewers verifying specialist work. For methodological principles, see Sections 3.4.3 (QED), 3.4.4 (RCT), and 3.4.4 (Realist Evaluation).

B.8.1 Quasi-Experimental Design (QED) Commissioning Specification

B.8.1.1 QED Requirements and Feasibility Considerations

What QED requires:

Methodological requirements:

- Comparison group demonstrably similar to participant group on observable characteristics

- Statistical methods addressing pre-existing differences between participants and comparisons (propensity score matching, difference-in-differences, regression adjustment, instrumental variables, regression discontinuity)
- Common support assessment (sufficient overlap in characteristics between groups enabling matching)
- Balance testing demonstrating groups comparable after matching/adjustment on measured covariates

Data requirements:

- Baseline data for both participants and potential comparisons enabling matching (demographics, baseline outcomes, selection factors)
- Identical outcome measurement for participants and comparisons (same instruments, same timing, same procedures)
- Sufficient follow-up data with acceptable attrition rates (<30% in both groups, similar across groups)

Sample size requirements:

- Minimum 200 participants + 200 comparisons for adequately powered QED detecting meaningful effect sizes
- Larger samples required if: expected effect sizes small, outcome measures have high variance, substantial missing data anticipated

QED Critical Feasibility Factors

Factor	Assessment Questions	Validity Threats if Inadequate
Comparison group plausibility	Can non-participants similar to participants be identified? What sources exist?	No plausible comparison → selection bias uncontrolled → causal claims unjustified
Baseline data availability	Do pre-intervention data exist enabling matching?	No baseline data → cannot demonstrate groups comparable → residual confounding likely
Sample size adequacy	Are sufficient participants and comparisons available?	Inadequate sample → low statistical power → false negatives, imprecise estimates
Outcome measurement equivalence	Can identical outcomes be measured for both groups?	Different measurement → observed differences may reflect measurement not programme effects

Validity threats in QED:

- **Selection bias:** Participants differ systematically from non-participants on unmeasured characteristics affecting outcomes—matching controls measured differences only
- **Confounding:** Unmeasured factors correlated with both participation and outcomes produce spurious associations
- **Regression to the mean:** Participants selected at extreme values naturally move toward average regardless of programme
- **Differential attrition:** Loss to follow-up differs between groups undermining initial comparability
- **Specification sensitivity:** Results depend heavily on matching method, covariates included, functional form—requires sensitivity analysis

QED provides strong causal evidence when requirements met but does not eliminate all threats to causal inference. Results represent best available causal estimate given constraints, not definitive proof of causation.

B.8.1.2 Scope Definition Requirements

Commissioners must specify to evaluators:

QED Scope Definition Requirements

Specification	What to Define
Research question	Clear causal question (e.g., "Does programme increase employment rates compared to what would have happened without programme?")
Population	Target population, eligibility criteria, sample size (participants and comparisons)
Comparison group source	Where will comparisons come from? (waiting list, geographic comparison, administrative data, comparison sites)
Primary outcome	Single primary outcome driving sample size calculation, measurement instrument, timing
Secondary outcomes	2-4 secondary outcomes (exploratory, not driving sample size)
Time horizon	Follow-up duration (6 months, 12 months, 24 months post-programme)
Available data	Baseline data availability, administrative data access, Data Sharing Agreements status
Budget and timeline	Evaluation budget, timeline (typically 12-18 months from design to final report), deliverables

Additional specifications:

- GDPR compliance and ethics approval: Arranged by commissioner or evaluator responsibility?
- Comparison group recruitment: Commissioner facilitates access or evaluator manages?
- Administrative data linkage: Which agencies involved, who negotiates access?

B.8.1.3 Quality Standards Checklist

Commissioners should verify:

Design Appropriateness: QED design appropriate given comparison group availability and data constraints

Research question clearly causal (not merely descriptive or associational)

Comparison group source justified with plausible equivalence rationale

Sample and Power: Sample size adequate for detecting meaningful effect sizes (power calculation documented)

Minimum 200 participants + 200 comparisons (more if effect sizes expected to be small)

Attrition rates acceptable (<30% participant group, <30% comparison group)

Attrition similar across groups (differential attrition threatens validity)

Comparison Group Quality: Comparison group source clearly documented

Comparisons drawn from similar population as participants

Selection into participation accounted for through matching or statistical control

Baseline equivalence demonstrated on measured covariates (balance tables provided)

Common support assessed (sufficient overlap in propensity score distributions)

Outcome Measurement: Primary outcome clearly specified a priori (not selected post-hoc based on what showed effects)

Outcomes measured identically for participants and comparisons (same instruments, same timing)

Validated instruments used where available

Follow-up timing appropriate for outcomes to emerge

Outcome assessors blinded to group assignment where feasible

Statistical Analysis: Matching or statistical adjustment method appropriate (propensity score matching, difference-in-differences, regression adjustment, instrumental variables, regression discontinuity)

Covariates included in matching/adjustment justified theoretically



- Sensitivity analysis conducted testing robustness to alternative specifications
- Multiple testing addressed if multiple outcomes (Bonferroni correction or pre-specified primary outcome)
- Effect sizes reported (not just p-values)—practical significance assessed

Causal Inference Validity: Threats to internal validity acknowledged (selection bias, confounding, regression to mean, attrition)

- Assumptions stated explicitly (e.g., unconfoundedness assumption for matching methods)
- Sensitivity to unobserved confounding assessed where feasible
- Alternative explanations considered
- Claims proportionate to design strength (QED provides strong evidence but not definitive proof)

Transparency and Documentation: Methods documented enabling replication

- Data sources specified
- Analytical code available or analysis steps described in detail
- Limitations acknowledged explicitly
- Pre-registration or analysis plan documented (prevents p-hacking)

B.8.2 Randomised Controlled Trial (RCT) Commissioning Specification

B.8.2.1 RCT Requirements and Feasibility Considerations

What RCT requires:

Methodological requirements:

- Random assignment to treatment or control conditions eliminating selection bias
- Allocation concealment preventing prediction of assignments (maintains randomisation integrity)
- Intention-to-treat analysis (participants analysed as randomized regardless of intervention receipt)
- Adequate statistical power for detecting clinically or practically meaningful effects (typically 80% power minimum)
- Standardised intervention protocol ensuring consistent implementation
- Outcome assessment blinded to treatment assignment where feasible

Ethical requirements:

- Equipoise (genuine uncertainty whether intervention helps—no clear evidence of benefit or harm)
- Research Ethics Committee approval before recruitment begins

- Informed consent protecting participant autonomy
- Acceptable control condition (standard care, delayed intervention, placebo where appropriate, or nothing if no standard exists)
- Data monitoring addressing participant safety during trial
- Trial registration (ClinicalTrials.gov, ISRCTN) before first participant enrolled

Sample size requirements:

- Minimum 300 individuals (150 per arm) for simple two-arm trial detecting moderate effects
- Substantially larger samples required if: effect sizes expected to be small, cluster randomisation used, multiple study arms, high attrition anticipated
- Sample size calculation based on primary outcome, expected effect size, variance, desired power, and significance level

Critical feasibility factors:

RCT Critical Feasibility Factors

Factor	Assessment Questions	Validity Threats if Inadequate
Ethical equipoise	Is there genuine uncertainty whether programme helps? Is randomisation ethically justifiable?	Lack of equipoise → Research Ethics Committee refuses approval → trial cannot proceed
Adequate sample	Are sufficient eligible individuals accessible within reasonable timeframe?	Inadequate sample → underpowered trial → uninformative results, wasted resources
Randomisation acceptance	Will stakeholders accept random assignment? Can implementation team maintain allocation concealment?	Stakeholder resistance → trial contamination, protocol violations → compromised validity
Control condition ethics	What happens to control group? Is this acceptable given equipoise and programme context?	Unacceptable control → ethics approval denied OR differential attrition undermining randomisation
Implementation stability	Is intervention standardised and consistently deliverable?	Unstable intervention → implementation variation undermines internal validity, cannot interpret results

Contexts typically using RCT:

- Government-commissioned evaluations with dedicated evaluation budgets
- Academic-practice partnerships where university provides trial infrastructure and expertise
- Large multi-site programmes with sufficient scale for adequate power
- Innovation testing where funders require experimental evidence
- EU research projects (HORIZON) where impact evaluation is grant requirement

Validity threats in RCT:

- **Inadequate allocation concealment:** Selection bias enters if assignments predictable
- **High attrition:** Loss to follow-up undermines randomisation if differential between groups
- **Implementation infidelity:** Intervention delivered inconsistently produces diluted effects
- **Contamination:** Control group accesses intervention elements diluting treatment contrast
- **External validity limitations:** RCT establishes efficacy under trial conditions but effectiveness in routine practice may differ

RCT represents gold standard for causal inference when requirements met. However, substantial infrastructure, resources, and ethical prerequisites mean most social service programmes cannot meet RCT requirements. Rigorous QED or transparent outcome monitoring often more appropriate given organisational capacity and programme context.

B.8.2.2 Scope Definition Requirements

Commissioners must specify to evaluators:

Scope Definition Requirements

Specification	What to Define
Research question	Clear efficacy question (e.g., "Does programme increase employment compared to standard Job Centre services?")
Population	Target population, inclusion/exclusion criteria, recruitment settings
Intervention	Standardised intervention protocol, intervention duration, implementation fidelity monitoring approach

Control condition	What control group receives (standard care, delayed intervention, placebo, nothing), rationale for control choice
Primary outcome	Single primary outcome driving sample size and inferential conclusions, measurement instrument (validated), timing
Secondary outcomes	2-5 secondary outcomes (exploratory), process measures, mechanism tests
Randomisation approach	Individual randomisation or cluster randomisation (sites, groups), randomisation ratio (1:1 typical, 2:1 if recruitment challenge)
Sample size	Target sample (from power calculation), expected attrition (plan for 20-30% loss to follow-up), over-recruitment strategy
Timeline	Trial duration (typically 18-36 months: design 3-6 months, recruitment 6-12 months, intervention 6-12 months, follow-up 3-6 months, analysis 3-6 months)
Budget	Comprehensive budget covering: trial infrastructure, research staff, participant compensation, data collection, intervention delivery, analysis, reporting

Mandatory requirements:

Ethics approval: Research Ethics Committee approval BEFORE recruitment

Trial registration: Prospective registration (ClinicalTrials.gov or ISRCTN) BEFORE first participant enrolled

CONSORT reporting: Commitment to follow CONSORT 2010 standards for reporting

B.8.2.3 Quality Standards Checklist

Commissioners should verify:

- Design and Ethics:** Ethical equipoise justified explicitly
 Research Ethics Committee approval obtained
 Trial prospectively registered (ClinicalTrials.gov, ISRCTN)
 Informed consent procedures protect participant rights
 Data monitoring plan addresses participant safety

- Randomisation:** Allocation concealment properly implemented (prevents selection bias)
 Randomisation sequence adequately generated (computer-generated,

unpredictable)

- Allocation conducted by independent party (not recruitment staff)
- Randomisation log maintained documenting every allocation
- Protocol violations documented with corrective actions

Sample Size and Power: Power calculation justifies sample size (80% power minimum to detect clinically meaningful effects)

- Sample size accounts for expected attrition (typically add 20-30%)
- Recruitment plan realistic achieving target sample within timeline

Outcome Measurement: Primary outcome pre-specified in trial registration

- Validated instruments used where available
- Outcome assessment blinded to treatment allocation where feasible (assessor unaware which group participants assigned)
- Follow-up timing appropriate for outcomes to emerge
- Missing data minimised (retention strategies, multiple contact methods)

Statistical Analysis: Analysis plan pre-specified (prevents p-hacking)

- Intention-to-treat analysis conducted (participants analysed as randomized regardless of intervention receipt)
- Per-protocol analysis secondary (complier-only analysis)
- Missing data handled appropriately (multiple imputation or sensitivity analysis)
- Subgroup analyses pre-specified (prevents fishing expeditions)
- Effect sizes and confidence intervals reported (not just p-values)

Implementation Fidelity: Intervention protocol standardised and documented

- Fidelity monitoring demonstrates consistent implementation
- Contamination assessed (did control group access intervention elements?)
- Process evaluation documents actual implementation (not just intended)

Reporting: CONSORT flow diagram documents participant progress

- Baseline characteristics table shows groups balanced after randomisation
- Attrition documented with reasons
- Adverse events reported (negative effects, harms)
- Limitations acknowledged
- Results reported regardless of statistical significance (prevents publication bias)

B.8.3 Realist Evaluation Commissioning Specification

B.8.3.1 Realist Evaluation Requirements and Contexts

What Realist Evaluation requires:

Methodological requirements:

- Explicit programme theory specifying Context-Mechanism-Outcome (CMO) configurations
- Theory-driven data collection testing how mechanisms work differently in different contexts
- Iterative theory refinement based on evidence (initial theory tested, revised, retested)
- Demonstration of context-mechanism interaction (not just that outcomes vary, but HOW mechanisms work differently in different contexts)
- Multiple cases or sites capturing contextual variation
- Mixed methods enabling theory testing (qualitative understanding of mechanisms, quantitative outcome patterns)

Data requirements:

- Purposive sampling capturing contextual variation (different implementation settings, participant characteristics, policy environments)
- Data sources enabling mechanism identification (interviews explaining HOW change happened, observations of programme operation, documents revealing programme theory)
- Sufficient depth for saturation or rich case understanding (typically 20-60 interviews, multiple sites/cases)
- Quantitative outcome data showing patterns across contexts (when available)

Expertise requirements:

- Skilled qualitative researcher with realist evaluation experience
- Theory-driven evaluation expertise (not just descriptive qualitative research)
- Understanding of mechanism identification and context-sensitive explanations

Resource implications:

- Timeline: Typically 12-24 months (theory development 2-3 months, data collection 6-12 months, iterative analysis 3-6 months, reporting 1-3 months)
- Budget: €30K-€150K depending on scale, data collection intensity, number of sites
- Intensive qualitative work: Interview transcription, analysis, theory refinement cycles

Contexts where Realist Evaluation particularly appropriate:

- Complex interventions with multiple interacting components operating through different mechanisms
- Programmes where outcomes depend substantially on context (implementation quality, participant readiness, local resources, policy environment)
- Understanding HOW and WHY programmes work valued alongside WHETHER they work
- Multi-actor environments where attribution to single programme impossible (contribution analysis integrated)
- Scaling or adaptation decisions requiring understanding which mechanisms essential and which contexts enabling
- Funders value qualitative depth and contextual understanding alongside quantitative patterns

Contexts where simpler approaches sufficient:

- Simple interventions with clear mechanisms operating similarly across standard contexts
- Definitive yes/no causal conclusion primary need (RCT/QED more appropriate)
- Programmes operating independently in controlled settings (realist complexity methods unnecessary)
- Limited resources precluding quality qualitative research investment

Validity threats when requirements not met:

- Inadequate theory development → descriptive not explanatory evaluation
- Insufficient contextual variation sampled → cannot demonstrate context-mechanism interaction
- Poor qualitative methods → superficial understanding, mechanisms not genuinely identified
- No theory refinement → initial theory imposed on data rather than tested and revised
- Over-claiming causation → contribution claims become attribution claims inappropriate for complex multi-actor settings

B.8.3.2 Scope Definition Requirements

Commissioners must specify:

Scope Definition Requirements

Specification	What to Define
Evaluation focus	What do you want to understand? "What works for whom in what circumstances?" or "How does programme create change?" or "Why does programme succeed here but not there?"
Initial programme theory	Starting Theory of Change or programme theory specifying: In what contexts (C), which mechanisms fire (M), producing which outcomes (O) for which participants
Contexts of interest	Which contextual variations to examine? (implementation quality, participant readiness, local resources, policy environment, cultural factors)
Data collection scope	How many sites/cases? What data sources? (interviews, observations, documents, quantitative outcome patterns)
Stakeholders to engage	Participants, staff, managers, partners, policymakers—who can explain how programme works?
Timeline	Typically 12-24 months (theory development 2-3 months, data collection 6-12 months, analysis 3-6 months, reporting 1-3 months)
Budget	Varies by scale and data collection intensity—obtain estimates from qualitative researchers with realist expertise

B.8.3.3 Quality Standards Checklist

Commissioners should verify:

Theory-Driven Approach:

- Initial programme theory explicit (CMO configurations specified clearly)
- Theory grounded in evidence (prior research, stakeholder knowledge, programme documentation)
- Theory testable through data collection (generates specific predictions about context-mechanism-outcome patterns)
- Theory refined iteratively based on evidence (not fixed from start)

Context-Mechanism-Outcome Configurations: Contexts specified (what implementation/participant/environmental factors matter?)

Mechanisms identified (underlying processes producing change—not activities themselves but how activities trigger change through reasoning, resources, relationships)

- Outcomes linked to specific CMO patterns
- CMO configurations demonstrate HOW mechanisms work differently in different contexts

Data Collection: Data sources adequate for testing theory (interviews, observations, documents, quantitative patterns)

- Purposive sampling captures contextual variation
- Sample size adequate for saturation or depth (typically 20-60 interviews, multiple sites/cases)
- Qualitative methods rigorous (interview protocols, systematic observation, document analysis)

Analysis: CMO configurations tested against data systematically

- Rival explanations considered (alternative theories explaining observed patterns)
- Theory refined based on evidence (contradictory cases lead to theory revision)
- Context-mechanism interaction demonstrated (shows HOW mechanisms work differently in different contexts, not just that outcomes vary)

Explanatory Power: Evaluation explains outcome variation (why programme succeeds for some participants/contexts but not others)

- Mechanisms made visible (not just "programme works" but "programme triggers X mechanism when Y context present producing Z outcome")
- Transferability assessed (which contexts enable programme success elsewhere? which require adaptation?)

Transparency: Theory evolution documented (initial theory, refinements, final theory)

- Data collection and analysis procedures described
- Evidence supporting CMO configurations presented
- Limitations acknowledged
- Sufficient detail enabling assessment of credibility

B.9: EU Integration and Cross-Border Evaluation Templates

Purpose: Coordinate evaluation across multiple EU member states with different languages, administrative systems, and policy contexts.

When to Use: EU-funded consortia involving 3+ countries; cross-border service delivery requiring comparable evaluation; need for both EU-level synthesis and national reporting; GDPR compliance across jurisdictions.

Key Challenges:

- Administrative data: Different systems, definitions, access rules across countries
- Language and culture: Translation ≠ equivalence; concepts may not transfer
- Data protection: Each country implements GDPR differently
- Quality control: Ensuring consistency across dispersed teams
- Reporting: Balancing EU synthesis with national specificity

B.9.1 Core Indicator Selection Framework

SELECTION PROCESS

Step 1: Propose indicators based on EU priorities and programme logic (Theory of Change)

Step 2: Assess feasibility in EACH partner country using criteria below

Step 3: Retain indicators feasible in all countries OR make optional with documentation

Step 4: Create common data dictionary with country-specific implementation notes

Template 19 - Cross-Border Evaluation Feasibility Assessment by Country

Proposed Indicator	Definition	Country A Feasible?	Country B Feasible?	Country C Feasible?	Decision	Notes
Employment rate at 6 months	% participants in employment (≥15 hrs/week) 6 months post-programme	<input type="checkbox"/> Yes <input type="checkbox"/> No Source: _____	<input type="checkbox"/> Yes <input type="checkbox"/> No Source: _____	<input type="checkbox"/> Yes <input type="checkbox"/> No Source: _____	<input type="checkbox"/> CORE (all) <input type="checkbox"/> OPTIONAL <input type="checkbox"/> DROP	
Education progression	% participants enrolling in further education within 12 months	<input type="checkbox"/> Yes <input type="checkbox"/> No Source: _____	<input type="checkbox"/> Yes <input type="checkbox"/> No Source: _____	<input type="checkbox"/> Yes <input type="checkbox"/> No Source: _____	<input type="checkbox"/> CORE <input type="checkbox"/> OPTIONAL <input type="checkbox"/> DROP	
[Add indicators]						

Feasibility Criteria (assess for each country)

Criterion	Assessment Question
Data availability	Does data source exist? (Admin records, routine collection, or feasible primary data collection)
Comparable definition	Is concept defined consistently? (e.g., "employment" - paid work, hours threshold, self-employment)
Administrative access	Can partner legally and practically access data? (GDPR, data sharing agreements, technical capacity)
Timing alignment	Can data be collected at required timepoints across all countries?
Cultural validity	Is concept meaningful in local context? (Not just translated, but culturally equivalent)
Cost-effectiveness	Is collection cost justified by multi-country value? (Some indicators expensive in certain countries)

INDICATOR TYPES

- CORE indicators: Collected by all partners (enables cross-country comparison)
- OPTIONAL indicators: Collected by subset of partners (report which partners collected)
- COUNTRY-SPECIFIC indicators: Address national priorities (reported separately)

HARMONISATION STANDARDS

Align with established frameworks where possible:

- Education: ISCED (International Standard Classification of Education)
- Occupation: ISCO (International Standard Classification of Occupations)
- Socioeconomic status: EU-SILC categories
- Health: ICD-10 (International Classification of Diseases)
- Migration: Eurostat definitions

B.9.2 Translation and Cultural Adaptation Protocol

Standard Translation Process (for surveys, interview guides)

Step	Process	Responsible	Timeline
1. Forward translation	Professional translator (English → target language)	Partner + certified translator	2 weeks
2. Expert review	Bilingual expert reviews for accuracy and cultural appropriateness	Subject matter expert	1 week
3. Back translation	Independent translator (target → English)	Different certified translator	2 weeks
4. Comparison	Compare original and back-translation; identify discrepancies	Lead evaluator + partner	1 week
5. Adjudication	Resolve discrepancies through consensus (whole team)	Evaluation working group	1 week
6. Cognitive testing	Test with 5-10 target population ("What does this mean?"); revise	Partner	2 weeks
7. Finalization	Document final version and all adaptations	Lead evaluator	1 week

Total timeline: Translations typically takes ~10 weeks per language (budget accordingly for multi-country projects)

Template 20 – Cultural Adaptation Documentation

Item Number	Original (English)	Language	Adapted Version	Rationale
Q5	"How confident are you managing your finances?"	Greek	[Greek text]	Cultural context: "managing finances" required framing around family budgeting practices specific to Greek context
Q12 example	"Did you receive benefits?"	Spanish	[Spanish text]	Term "benefits" ambiguous - specified "social welfare benefits (desempleo, ayudas)"

CRITICAL PRINCIPLE: Aim for conceptual equivalence (same meaning) not literal translation (same words). Example: "Job security" literal translation may not capture actual meaning in precarious labour markets.

Cultural Adaptation Considerations

Element	Issue	Solution
Response scales	Some cultures avoid extremes (e.g., 1 or 7 on 7-point scale)	Use verbal labels for each point; analyse culturally-specific response patterns
Examples	Examples meaningful in one culture may not transfer	Develop country-specific examples maintaining same construct
Sensitive topics	Topics taboo or sensitive differ by culture (income, mental health)	Adapt wording, offer skip options, ensure confidentiality emphasized
Formality	Languages differ in formality requirements (tu/vous in French)	Match formality to cultural norms and target population

B.9.3 Data Harmonisation Procedures

DATA HARMONISATION WORKFLOW

Step 1: Each partner collects data using adapted instruments

Step 2: Partners submit data in agreed format (see data dictionary)

Step 3: Lead evaluator conducts quality checks (Step 4 below)

Step 4: Variables harmonised according to protocol (Step 5 below)

Step 5: Create master dataset with country identifiers

Step 6: Document harmonisation decisions and country-specific notes

Common Harmonisation Scenarios

Challenge	Example	Harmonisation Rule	Implementation
Different categories	Education systems vary: Country A: 5 levels Country B: 8 levels Country C: 4 levels	Map to ISCED standard	Create mapping table for each country: Country A Level 3 = ISCED 3 Country B Levels 4-5 = ISCED 3 etc.
Different scales	Income: Country A: € Country B: £ Country C: SEK	Convert to € using: 1. Exchange rate (date: ____) 2. PPP adjustment (Eurostat)	Document conversion: £25K × 1.17 × 0.88 (PPP) = €25,740
Different timing	Programme length: Country A: 6 months Country B: 12 months Country C: 9 months	Standardize measurement points: T0 = Baseline T1 = 6 months T2 = 12 months	All countries measure at T0, T1, T2 regardless of programme length
Different definitions	"Employment": Country A: any paid work Country B: ≥15 hrs/week Country C: formal contract only	Agree common definition upfront	Use most restrictive common definition: ≥15 hrs/week with contract OR self-employment registered

Data Quality Checks

Check Type	What to Check	Action if Problem
Completeness	Missing data rates per country (>30% concern)	Investigate reason; assess bias; document
Range	Values within expected range (outliers, impossible values)	Query with partner; correct or code as missing
Consistency	Internal consistency (e.g., age vs birth year)	Query with partner; resolve
Distribution	Distributions similar across countries (or explainable differences)	If unexpected: investigate measurement issue vs real difference
Response patterns	Unusual patterns (all same response, systematic missing)	Investigate data collection quality

Template 21 – Data Dictionary Template

Variable Name	Label	Type	Coding	Harmonisation	Notes
country	Country	Categorical	1=Greece, 2=Spain, 3=Austria, etc.	N/A	
age	Age in years	Continuous	Range: 18-65	Calculated from birth year if needed	

educ_isc	Education level	Categorical	0=Pre-primary, 1=Primary, 2=Lower secondary, 3=Upper secondary, 4=Post-secondary non-tertiary, 5=Short-cycle tertiary, 6=Bachelor, 7=Master, 8=Doctoral	Mapped from national systems - see country mapping tables	Greece: Mapping table A.1; Spain: Mapping table A.2
employ_t1	Employed at T1	Binary	0=No, 1=Yes	Definition: ≥15 hrs/week OR registered self-employment	Spain: Includes "economía sumergida" if confirmed by participant

B.9.4 Multi-Level Reporting Templates

Reporting Architecture

Level	Audience	Content	Length	Language
EU Consortium Report	EC Project Officer, all partners, public	Cross-country synthesis, comparative analysis, EU policy implications	50-80 pages + appendices	English
National Reports	National funders, ministries, stakeholders	Country-specific findings in context, national policy implications, comparison to other countries	25-40 pages	National language + English summary

Partner/ Local Briefs	Local managers, practitioners, participants	Site-specific results, practical recommendations, accessible format	5-10 pages	National language
----------------------------------	---	---	------------	-------------------

EU CONSORTIUM REPORT STRUCTURE

1. Executive Summary (3 pages)

- Cross-country key findings
- Recommendations for EU policy
- Country-specific highlights

2. Introduction (5 pages)

- Programme overview and EU context
- Evaluation questions and design
- Partner countries and sites

3. Methods (8 pages)

- Cross-border evaluation design
- Data collection and harmonisation
- Translation and adaptation
- Analysis approach (aggregated, comparative, context-mechanism)
- Limitations

4. Cross-Country Synthesis (20-30 pages)

- Aggregated findings (pooled analysis across all countries)
- Comparative analysis (country-by-country comparison tables)
- Pattern identification (which contexts produce which outcomes?)
- Implementation variation and fidelity

5. Country Summaries (2-3 pages each)

- Brief overview of findings per country
- Contextual factors
- Unique features or challenges

6. Discussion (10 pages)

- Interpretation of cross-country patterns
- Transferability and scaling considerations
- EU policy implications
- Recommendations for practice

7. Appendices

- Country-specific detailed results
- Technical methods details
- Data collection instruments

Comparative Analysis Approaches

Approach	Method	When to Use	Example Presentation
Aggregated	Pool all countries (N=total), report overall effect	For overall EU-level conclusion; testing programme model	Table: Outcome All Countries Mean 95% CI p-value
Country Comparison	Present results by country, compare	Understanding variation; identifying best practices	Table: Outcome Country A Country B Country C Pattern
Context-Mechanism	Link outcomes to contextual factors across countries	Understanding "what works for whom in what circumstances"	Table: Context Countries with Context Outcome Mechanism

Recommendation: Use ALL THREE approaches for comprehensive understanding. Aggregated shows overall effect, country comparison shows variation, context-mechanism explains why.

B.9.5 Cross-Border Coordination Checklist

Template 22 - Cross-Border Partnership Governance

Element	Status
Lead evaluator designated	<input type="checkbox"/> Complete (Partner: ____ Person: ____)
Evaluation working group established	<input type="checkbox"/> Complete (Members from each partner)
Partner responsibilities defined in consortium agreement	<input type="checkbox"/> Complete (Signed: ____)
Budget allocated per partner for evaluation activities	<input type="checkbox"/> Complete (Documented)
Coordination meeting schedule	<input type="checkbox"/> Complete (Monthly <input type="checkbox"/> Quarterly <input type="checkbox"/> Other: ____)
Communication platform established	<input type="checkbox"/> Complete (Platform: ____ Access: ____)
Document sharing system	<input type="checkbox"/> Complete (Secure, GDPR-compliant: ____)
Decision-making protocol	<input type="checkbox"/> Complete (Consensus <input type="checkbox"/> Lead decides <input type="checkbox"/> Voting)

Template 23 - Cross-Border Data Protection & Ethics Coordination

Element	Status
Common data protection protocol agreed	<input type="checkbox"/> Complete (see Appendix A.4 for EU data transfer rules)
Data sharing agreements signed	<input type="checkbox"/> Complete (All partners signed: ____)
Ethics approvals obtained	<input type="checkbox"/> Country A <input type="checkbox"/> Country B <input type="checkbox"/> Country C [add all]
Consent forms harmonized	<input type="checkbox"/> Complete (Common core + country adaptations)
Data security protocols implemented	<input type="checkbox"/> Complete (Encryption, access control, storage)
GDPR compliance documented	<input type="checkbox"/> Complete (Each partner DPO involved)

Template 24 – Cross-Border Quality Assurance Coordination

Activity	Implementation
Data collection training	<input type="checkbox"/> Centralized (all partners together) <input type="checkbox"/> Materials shared (partners train locally) Date: ____ Participants: ____
Pilot testing	<input type="checkbox"/> Complete in all countries (Date: ____) <input type="checkbox"/> Issues resolved (Documented: ____)
Data quality monitoring	<input type="checkbox"/> Partners submit quality reports (Frequency: ____) <input type="checkbox"/> Lead conducts quality checks <input type="checkbox"/> Issues documented and resolved
Analysis coordination	<input type="checkbox"/> Shared analysis plan <input type="checkbox"/> Common software/syntax shared <input type="checkbox"/> Joint analysis meetings scheduled
Report drafting	<input type="checkbox"/> Templates agreed <input type="checkbox"/> Partners contribute sections <input type="checkbox"/> Lead synthesizes <input type="checkbox"/> Partners review drafts

B.9.6 Cross-Border Evaluation Quality Assurance
QUALITY CHECKLIST
Design Quality:

- Evaluation design appropriate for cross-country comparison
- Core indicators feasible in all partner countries (assessed using B.8.1)
- Country-specific adaptations documented and justified
- Methods appropriate for multi-site implementation

Translation & Adaptation Quality:

- Full translation protocol followed (forward, back, review, cognitive testing)
- Adaptations documented with rationale
- Conceptual equivalence achieved (not just literal translation)
- Pilot tested in each language

Data Collection Quality:

- Data collectors trained to common standard (or using shared materials)

- Data collection monitored across all sites
- Response rates adequate in all countries (document if disparate)
- Data quality checks conducted (see B.8.3)

Harmonisation Quality:

- Variables harmonised according to protocol (B.8.3)
- Data dictionary complete and shared with all partners
- Country-specific issues documented
- Comparability verified (distributions, patterns reasonable across countries)

Analysis Quality:

- Methods appropriate for multi-country data (e.g., multilevel/mixed-effects models)
- Country as clustering variable considered
- Country-level variations reported (not just aggregated)
- Contextual factors linked to outcomes
- Aggregation appropriate (not masking important country differences)

Partnership Coordination:

- All partners contributed to evaluation design
- Regular coordination maintained throughout (meeting minutes documented)
- Data shared according to agreements and timelines
- Quality issues identified and addressed collaboratively
- Partners reviewed and approved reports

Reporting Quality:

- Multilevel reporting implemented (EU, national, local)
- Cross-country synthesis clear and well-supported
- Country-specific results presented with context
- Transferability discussed (what works where and why)
- Limitations acknowledged (translation, harmonisation, country differences)
- EU policy implications articulated

CRITICAL ISSUES:

- Core indicators not feasible in some countries (comparability compromised)
- Translation inadequate (cultural invalidity risk)
- Data protection coordination inadequate (legal/ethical risk)
- Partners not coordinating effectively (inconsistent implementation)
- Quality monitoring absent (cannot trust cross-country comparisons)
- Country differences ignored in analysis (aggregation hides important variation)

Bibliography

Ackerman, F., & Heinzerling, L. (2004). *Priceless: On knowing the price of everything and the value of nothing*. The New Press.

Anderson, A. A. (2005). *The community builder's approach to theory of change: A practical guide to theory development*. Aspen Institute Roundtable on Community Change.

Belton, V., & Stewart, T. J. (2002). *Multiple criteria decision analysis: An integrated approach*. Kluwer Academic Publishers. DOI: [10.1007/978-1-4615-1495-4](https://doi.org/10.1007/978-1-4615-1495-4)

Boardman, A. E., Greenberg, D. H., Vining, A. R., & Weimer, D. L. (2018). *Cost-benefit analysis: Concepts and practice* (5th ed.). Cambridge University Press.

Business in the Community. (2012). *Social return on investment of Ready for Work*.

Centre for Social Justice. (2021). *Housing First: Housing led solutions to rough sleeping and homelessness*. <https://www.centreforsocialjustice.org.uk/library/housing-first-housing-led-solutions-to-rough-sleeping-and-homelessness>

Christensen, T. N., Kruse, M., Hellström, L., & Eplov, L. F. (2020). Cost-utility and cost-effectiveness of individual placement support and cognitive remediation in people with severe mental illness: Results from a randomized clinical trial. *European Psychiatry*, 64(1), e3. <https://doi.org/10.1192/j.eurpsy.2020.111>

Christensen, T. N., Wallstrøm, I. G., Stenager, E., Bojesen, A. B., Gluud, C., Nordentoft, M., & Eplov, L. F. (2019). Effects of individual placement and support supplemented with cognitive remediation and work-focused social skills training for people with severe mental illness: A randomized clinical trial. *JAMA Psychiatry*, 76(12), 1232–1240. <https://doi.org/10.1001/jamapsychiatry.2019.2291>

Christensen, T. N., Wallstrøm, I. G., Stenager, E., Hellström, L., Bojesen, A. B., Nordentoft, M., & Eplov, L. F. (2023). 30-month follow-up of individual placement and support (IPS) and cognitive remediation for people with severe mental illness: Results from a randomized clinical trial. *Psychiatry Journal*, 2023, 2789891. <https://doi.org/10.1155/2023/2789891>

Department for Work and Pensions. (2008). *Cost benefit analysis framework for employment programmes*. DWP. <https://assets.publishing.service.gov.uk/media/5a7cd3ebe5274a34d8d332f0/WP86.pdf>

Dodgson, J. S., Spackman, M., Pearman, A., & Phillips, L. D. (2009). *Multi-criteria analysis: A manual*. Department for Communities and Local Government.

Drake, R. E., Bond, G. R., & Becker, D. R. (2012). *Individual placement and support: An evidence-based approach to supported employment*. Oxford University Press.

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes* (4th ed.). Oxford University Press.

European Commission. (2021). *Better regulation toolbox*. https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation/better-regulation-guidelines-and-toolbox_en

European Commission. (2022). *One-stop-shop guidance centres for young people (Ohjaamo)*. European Social Fund Plus. <https://ec.europa.eu/european-social-fund-plus>



European Commission. (2023). Commission Delegated Regulation (EU) 2023/2772 supplementing Directive 2013/34/EU as regards sustainability reporting standards. *Official Journal of the European Union*. http://data.europa.eu/eli/reg_del/2023/2772/oj

European Commission. (2024). *Horizon Europe programme guide* (Version 3.0). https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

European Parliament, Council of the European Union, & European Commission. (2017). European Pillar of Social Rights. *Official Journal of the European Union*, C 428, 10–15. [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017C1213\(01\)](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017C1213(01))

European Social Fund Plus. (n.d.). *One-stop-shop guidance centres for young people (Ohjaamo)*. European Commission. <https://european-social-fund-plus.ec.europa.eu/en/social-innovation-match/case-study/one-stop-shop-guidance-centres-young-people-ohjaamo>

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

European Union. (2021). Regulation (EU) 2021/1057 of the European Parliament and of the Council of 24 June 2021 establishing the European Social Fund Plus (ESF+). *Official Journal of the European Union*, L 231, 21–75. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R1057>

Fujiwara, D., & Campbell, R. (2011). *Valuation techniques for social cost-benefit analysis*. HM Treasury & Department for Work and Pensions. <https://www.gov.uk/government/publications/valuation-techniques-for-social-cost-benefit-analysis>

Hatry, H. P., Cowan, J., Weiner, K. M., & Lampkin, L. M. (2006). *Outcome management for nonprofit organizations: A practical guide to getting results*. Urban Institute Press.

HM Treasury. (2022). *The Green Book: Central government guidance on appraisal and evaluation*. <https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government>

ICF Consulting. (2024). *Evaluation of the Housing First pilots: Cost benefit analysis — final report*. Ministry of Housing, Communities and Local Government. https://assets.publishing.service.gov.uk/media/671a6fea603993b7a8f75db5/Housing_First_Cost_Benefit_Analysis_Report.pdf

INNOSI. (2017). *Work package 4 case study report: Finland Youth Guarantee / Ohjaamo*. Horizon 2020 project, Association of Finnish Local and Regional Authorities.

Kautto, T., Korpilauri, T., Pudas, M., & Savonmäki, P. (2018). *One-stop guidance center (Ohjaamo)* (M. Määttä, Ed.). <http://www.doria.fi/handle/10024/162148>

Khan-Gökkaya, S., Higgen, S., & Mösko, M. (2019). Qualification programmes for immigrant health professionals: A systematic review. *PLoS ONE*, 14(11), e0224933. <https://doi.org/10.1371/journal.pone.0224933>

Khan-Gökkaya, S., & Mösko, M. (2020). Process- and outcome evaluation of an orientation programme for refugee health professionals. *Medical Education Online*, 25(1), 1811543. <https://doi.org/10.1080/10872981.2020.1811543>



Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels*. Berrett-Koehler.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>

Kubiszewski, I., Costanza, R., Eastoe, J., Lu, T., Mulder, K., Patteson Hernandez, G., Benczúr, P., & Dixon-Declève, S. (2025). Building consensus on societal wellbeing: A semantic synthesis of indicators to move beyond GDP. *Ecological Indicators*, 178, 114076. <https://doi.org/10.1016/j.ecolind.2025.114076>

Määttä, M. (2019). Reforming youth transition support with the multi-agency approach? A case study of the Finnish one-stop guidance centers. *Sociologija*, 61(2), 277–291. <https://doi.org/10.2298/SOC1902277M>

Mayne, J. (2012). Contribution analysis: Coming of age? *Evaluation*, 18(3), 270–280. <https://doi.org/10.1177/1356389012451663>

Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Beltz Verlag.

Ministry of Housing, Communities and Local Government. (2024). *Evaluation of the Housing First pilots: Final synthesis report*. https://assets.publishing.service.gov.uk/media/671a70221898d9be93f75db4/Housing_First_Final_Synthesis_Report.pdf

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O’Cathain, A., Tinati, T., Wight, D., & Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *BMJ*, 350, Article h1258. <https://doi.org/10.1136/bmj.h1258>

Nicholls, J., Lawlor, E., Neitzert, E., & Goodspeed, T. (2012). *A guide to social return on investment*. The SROI Network. <https://socialvalueint.org/social-value/sroi/>

Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Harvard University Press. <https://doi.org/10.4159/harvard.9780674061200>

Octavia Foundation. (2012). *Placing a value on work: A social return on investment report*. https://www.octaviafoundation.org.uk/assets/0000/1500/SROI_Report_Guardian_Version.pdf

OECD. (2019). *Investing in youth: Finland*. OECD Publishing. <https://doi.org/10.1787/1251a123-en>

OECD. (2021). *Applying evaluation criteria thoughtfully*. OECD Publishing. <https://doi.org/10.1787/543e84ed-en>

OECD DAC Network on Development Evaluation. (2019). *Better criteria for better evaluation: Revised evaluation criteria definitions and principles for use*. OECD Publishing. <https://www.oecd.org/dac/evaluation/revised-evaluation-criteria-dec-2019.pdf>

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. SAGE Publications.

Pleace, N., & Culhane, D. (2016). *Better than cure? Testing the case for enhancing prevention of single homelessness in England*. Crisis.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, Article c332. <https://doi.org/10.1136/bmj.c332>



Sen, A. (1999). *Development as freedom*. Oxford University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Social Value International. (2024). *The principles of social value*. <https://www.socialvalueint.org/principles>

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>

Stiglitz, J. E., Sen, A., & Fitoussi, J.-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Commission on the Measurement of Economic Performance and Social Progress.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., & Stewart-Brown, S. (2007). The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Development and UK validation. *Health and Quality of Life Outcomes*, 5, Article 63. <https://doi.org/10.1186/1477-7525-5-63>

Tzivanakis, N., Melios, G., & Moore, H. (2025). *Meta-analysis of existing indicators* (Deliverable D1.1). BENEFITS Project (Grant Agreement No. 101179032) – Horizon Europe.

Van den Eynden, V., Corti, L., Woollard, M., Bishop, L., & Horton, L. (2011). *Managing and sharing research data: A guide to good practice*. UK Data Archive, University of Essex. <https://dam.ukdataservice.ac.uk/media/622417/managingsharing.pdf>

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts* (pp. 65–92). Aspen Institute.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, Article 160018. <https://doi.org/10.1038/sdata.2016.18>

Wong, G., Westhorp, G., Manzano, A., Greenhalgh, J., Jagosh, J., & Greenhalgh, T. (2016). RAMESES II reporting standards for realist evaluations. *BMC Medicine*, 14, Article 96. <https://doi.org/10.1186/s12916-016-0643-1>



BENEFITS

Building Economic, Needs-Based and Environmental
evaluation Frameworks for Inclusive Transformation
of Social services in Europe



Funded by
the European Union