Using a Realist Framework to Overcome Evaluation Challenges in the Uncertain Landscape of Carbon Finance

Callum Murdoch, Lisa Keppler, Tillem Burlace, and Christine Wörlen

Abstract

In 2013, the United Kingdom Department for International Development and the Department of Business, Energy, and Industrial Strategy published a business case for the Carbon Market Finance Programme (CMFP). The core mandate: to build capacity and develop aids for least developed countries in sub-Saharan Africa to access finance via the carbon market. The chosen strategy involved signing emission reduction purchase agreements with private sector enterprises, using the United Nations Framework Convention for Climate Change's Clean Development Mechanism (CDM) to verify generation of tradeable certified emissions reductions. The World Bank's Carbon Initiative for Development (Ci-Dev) would implement the 12-year program. The team for the 2019 midterm evaluation found that program uncertainty—from sociopolitical challenges in pilot markets to global indecision on the future of Article 6 and carbon markets—would complicate assessing progress toward business case objectives. The collapse and failed recovery of the carbon market impacted underlying assumptions of the CMFP's theory of change, and uncertainty about CDM's future complicated evaluation of program sustainability. This chapter presents a practical approach to using realist evaluation to overcome the contextual uncertainties of the carbon market landscape, providing strengths and weaknesses of the approach applied and recommending a revised approach for future evaluations.

The Evaluation of the Carbon Market Finance Programme

In 2013, the UK Department for International Development (DFID) and the Department of Business, Energy, and Industrial Strategy (BEIS; then the Department of Energy and Climate Change) published a business case for the Carbon Market Finance Programme (CMFP) under the UK's International Climate Finance (ICF). The core mandate of the program was to build capacity and develop tools and methodologies to help least developed countries in sub-Saharan Africa access finance via the carbon market. The business case explored several options before settling on a strategy that involved signing emission reduction purchase agreements with private sector enterprises seeking to improve energy access

C. Murdoch (✉) · T. Burlace
NIRAS-LTS International, Edinburgh, UK
e-mail: callum-murdoch@ltsi.co.uk

L. Keppler · C. Wörlen
Arepo, Berlin, Germany

in least developed countries, using the United Nations Framework Convention for Climate Change's (UNFCCC) Clean Development Mechanism (CDM) to verify the generation of tradeable certified emissions reductions (CERs). The business case team selected the World Bank's Carbon Initiative for Development, or Ci-Dev, for the implementation of the program over a 12-year period.

On paper, the CMFP is a relatively straightforward results-based finance (RBF) program using carbon credits as the underlying result. However, the market for certified emissions reductions—and indeed, nearly all emission trading schemes—changed drastically since 2013 and was in a state of such uncertainty as of 2019 that market insiders were reticent to discuss it.[1] Starting in late 2011, the price for carbon trading instruments, including certified emissions reductions, began to decline. At the release of the CMFP business case, the market was in the middle of what most (including those at the UNFCCC) would describe as a collapse, and the outlook for recovery was difficult to determine (CDM Policy Dialogue, 2012). A further development that contributed to the uncertainty was the establishment of the Paris Agreement at the 21st Conference of Parties (COP21) in 2015, especially Article 6.4, which stated:

> A mechanism to contribute to the mitigation of greenhouse gas emissions and support sustainable development is hereby established under the authority and guidance of the Conference of the Parties serving as the meeting of the Parties to this Agreement for use by Parties on a voluntary basis. (UNFCCC, 2015)

The implication of this Article, and indeed the rest of Article 6, was that a new carbon verification and trading instrument would be established by the Paris Agreement, likely replacing the CDM. What Article 6 did not do was clearly establish what that mechanism will look like or how it will operate. The quest for the elusive Paris Rulebook, which would provide the foundations on which this new mechanism will be built, has not yet yielded any results (as of 2020). With the review of the nationally determined contributions scheduled for 2020, it was hoped that COP25 in 2019 would generate clarity. Although some progress was made, the future of carbon markets remains as undefined and uncertain as it has since the signing of the Paris Agreement.

The UK is implementing the CMFP in this context, accompanied by the evaluation described in this chapter. The 11-year evaluation kicked off in late 2014 and will run until the conclusion of the program in 2025. In 2019, the evaluation team conducted a midterm evaluation to gauge the program's progress to date (LTS International, 2020). We quickly found that the uncertainty surrounding the program, from the local sociopolitical challenges of the markets where the projects are being piloted to the global indecision on the future of Article 6 and carbon markets, would make assessing the program's progress toward its stated business case objectives a challenging process. The collapse and failed recovery of the market struck at many of the underlying assumptions of the CMFP's theory of change. Moreover, uncertainty about the future of CDM complicated any evaluative judgement on program sustainability.

The team considered a number of approaches before deciding that the realist evaluation approach (Pawson & Tilley, 1997) would be best suited to evaluating the program given the uncertain landscape. Although systematic evaluation methods are most common, they focus on explaining whether or not an intervention led to a certain outcome. Instead, realist evaluation helps to open the black box of program theory—it tries to explain how and why projects work or do not work, for whom, and under what circumstances. It recognizes that the context in which individual projects are operating makes important differences to the projects' results. It also shows that no project intervention is likely to work every-

---

[1] Even though the team for the evaluation discussed in this chapter noted this risk and limitation throughout the preparation of the evaluation approach, we were nonetheless surprised how unwilling critical senior stakeholders were to share information and opinions when the topic of post-Paris carbon markets was raised.

where, under all circumstances, and for everyone. The application of realist evaluation, therefore, provides the opportunity to evaluate projects within their unique and changing local, national, and global contextual factors.

Another benefit of realist evaluation is that it does not prescribe a specific, regimented approach. A plethora of literature offers different applications for realist principles in evaluation and various philosophical discussions about the nature of realist thinking. This chapter fits within the former category and seeks to offer a field guide of sorts for applying realist evaluation to interventions for which the underlying theory has been affected by the uncertain contextual landscape in which it operates. We first describe the methodology applied in the CMFP midterm evaluation, then weigh the positives and negatives of this approach, finally presenting a revised methodology taking into account the learning emerging from the evaluation.

Overview of Methodology

To systematically address the described complexity of both the program itself and of its embedded environment, the team developed an evaluation framework drawing on realist evaluation principles. Complementing this were other evaluative analysis methods, each of which gives specific insights into the program's dynamic implementation and progress. Descriptive analysis focused on verifiable and quantitative data, including reporting against the program's logical framework, a value-for-money analysis, benchmark assessments, and, to a lesser extent, a qualitative comparative analysis. We used explanatory analysis for qualitative, interpretative data and where the evaluation required greater consideration of the contextual factors contributing to the program's progress. For this explanatory analysis, the team used a realist evaluation approach in two tranches: first, as a specific evaluation method to gather, code, and analyze data from a variety of sources, and second, as a synthesis framework against which other explana-

tory evaluation methods, such as a contribution analysis and an energy market barrier analysis based on the Theory of No Change (Wörlen et al., 2011), could be assessed in the context of the wider portfolio findings. Figure 1 shows the chosen realist evaluation framework.

Realist Evaluation as an Approach

The leading questions of a realist evaluation ask how, why, for whom, and under what circumstances the program works or does not work. Answering these questions requires identifying the underlying generative mechanisms and causal relationships of the program's dynamic through a continuous, multistage hypotheses development process. This retroductive process moves back and forth between inductive and deductive logic based on assumptions, continuous learning, and the expertise of its developers (Greenhalgh et al., 2017a). Inductive reasoning generates a new theory from collected data and multiple observations showing a logical pattern, whereas deductive reasoning starts with a theory and the formulation of hypotheses which are tested and verified by observations. The retroductive approach applied by realist evaluation draws from both.

Hypothesis Development
The first step in the process was the formulation of hypotheses in the form of intervention-context-mechanism-outcome statements, or ICMOs, which were developed with a top-down approach. Based on the program's theory of change and a review of program- and context-related literature, these statements were formulated in an abstract way to be valid to the whole program itself, and to the project portfolio. ICMO configurations are the core analytical elements of realist evaluation. As shown in Fig. 2, they bring together in one statement

- a program's intervention (I)
- the context (C) in which the intervention takes place and that influences whether an intervention activates a mechanism (M), which is the
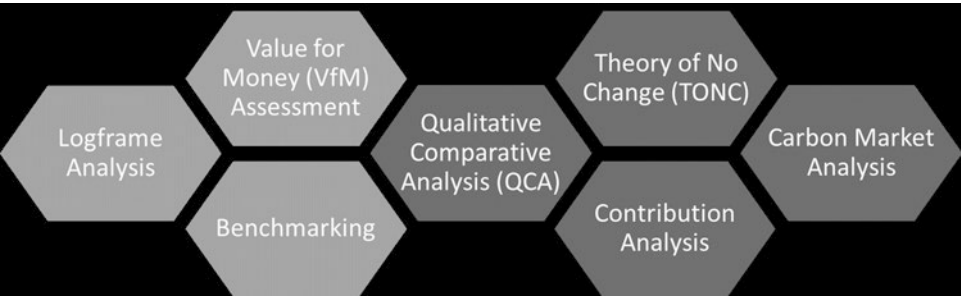
Fig. 1 Realist evaluation framework combining descriptive and explanatory evaluation methods

Fig. 2 Elements of an ICMO statement

response of the intervention target to the intervention

- the outcome (O), the desired end result of the other three components' interactions.

The intervention is the only factor under the direct control of the program; context, mechanism, and outcome are outside its direct control. The mechanism is the center of the realist explanation for how and why change occurs. It is a non-observable process, often described as changes in the reasoning and behavior of individuals or different levels of systems, that leads from the intervention to the outcome interconnected with contextual factors (Greenhalgh et al., 2017b).

Overall, our evaluation team developed four ICMO statements: two addressing the direct results of the program and two addressing the program's impact level, focusing on (a) barrier removal in energy markets and (b) the transformation of the carbon market, including replication of the program's approach (LTS International, 2020). An example ICMO statement (paraphrased from the CMFP evaluation) is provided in Box 1.

Box 1: Example ICMO Statement

By providing carbon-results-based financing and business development support funding (I), in a context with sufficient customer demand for the energy technologies, access to finance for the pilot enterprises, and a supportive policy framework (C), revenue and capacity for projects will be sufficient to overcome the operational challenges in providing rural energy access technologies (M), resulting in increased energy access and the generation and sale of certified emissions reductions (O).

To improve the explanatory value, each ICMO configuration consisted of sub-statements for each of the elements where the hypotheses involved compound statements (i.e., I1a, I1b, I1c, C1a, C1b, etc.; see Box 2). For example, the CMFP funding provided for carbon credit purchase often sought to trigger the same or similar mechanisms as funding provided for project readiness or capacity development. Thus, these two interventions were often considered in paral-

lel. However, a particular piece of evidence collected might support one intervention more than the other and therefore would need to be independently assessed.

Box 2: Example ICMO Sub-statements

By providing carbon-results-based financing (I1a) and business development support funding (I1b), in a context with sufficient customer demand for the energy technologies (C1a), access to finance for the pilot enterprises (C1b) and a supportive policy framework (C1c). . .

After splitting the ICMOs into sub-statements, the team identified critical components for each sub-statement that would prove the sub-statement's accuracy (see Box 3). The components listed were not exhaustive but served to guide what evidence would either confirm or disprove a sub-statement. This was to ensure consistency across users of the realist methodology and to enhance the deductive side of the analysis.

Box 3: Example ICMO Sub-statement Components

Sub-statement: "By providing carbon-results-based financing (I1a)".
　　Sample components:

1. An emissions reduction purchase agreement (ERPA) has been signed with the project entity.
2. The ERPA price is within a suitable range.
3. Finance has been transferred in exchange for carbon credits.

A fundamental element of realist evaluation is the involvement of core program or intervention stakeholders in the ICMO development process. In the CMFP evaluation, the ICMOs were refined through consultations with both the evaluation commissioners (BEIS) and the implementation agents (Ci-Dev) to ensure that they contained all of the relevant elements to depict the theory of change and its influences. The ICMOs passed through several iterations, often with minor phrasing or order changes.

Coding System

The evaluation team then tested the ICMO hypotheses by coding initial primary data (interviews) and secondary data (program documents, previous evaluation exercises, etc.) against the hypotheses to find out whether these theories were pertinent, productive, and appropriately designed. In an inductive process, we revised the hypotheses where the coded data gave indication about contexts, mechanisms, or outcomes that had not yet been considered in the ICMO statements, but where data showed relevance to the program's development. During the process, the team again consulted stakeholders on the interpretation of the available data to reach a reasonable judgement about the most useful findings. We repeated these steps at key stages of the evaluation process, leading to the retroductive nature of the approach.

For coding the ICMOs, we adapted the approach first adopted by the midterm evaluation of the UK Government's Climate Public Private Partnership Program (Climate Policy Initiative & LTS International, 2018). A matrix was designed that not only coded the evidence against the ICMO hypotheses, but also assessed the strength of each point of evidence and the overall evidence saturation. The intention of the coding matrix was to develop a quantifiable scoring process for the evidence collected that would allow the evaluation team to determine how accurate the initial hypotheses were. We devised a simple scoring system (see Table 1) that ranged from 3 (when a particular piece of evidence demonstrated high accuracy of the statement or sub-statement) to −3 (when evidence strongly disproved or contradicted the statement or sub-statement). The scoring also included a neutral value, X, which marked evidence as being relevant to the ICMO but not applicable to the current sub-statements. The team closely reviewed

Table 1 ICMO evidence scoring guide

Score	Description
3	Evidence strongly supports ICMO statement. Multiple or all components are met, or particularly strong evidence toward select components is provided.
1	Evidence partially supports ICMO statement. Some components have been met, or evidence supports the overall statement without actually meeting the component.
−1	Evidence partially contradicts ICMO statement. Evidence disproves or creates doubt that some components have been met, or evidence contradicts the overall statement without opposing specific components.
−3	Evidence strongly contradicts or disproves ICMO statement. Multiple or all components are countered or disproved, or particularly strong evidence negating select components is provided.
X	Evidence does not support or contradict the sub-statement but is relevant to the overall ICMO. For example, evidence provides a specific contextual factor that the statement hasn't captured, or an outcome that differs from those anticipated.

these evidence points and ensured the continuous verification of the relevance and revisions of the ICMO configurations. This scoring approach was designed to be simple, intuitive, and quantifiable while providing a traceable roadmap of how evidence was used to formulate specific evaluation findings.

To assess the strength of evidence, we categorized each piece of evidence according to Table 1, then used a modifier to weight the evidence (see Table 2). Verifiable evidence, either factual information or evidence from highly authoritative sources, received a two-times modifier to reflect the inherent strength of such evidence. Plausible evidence largely refers to data such as stakeholder interviews or discussions, qualitative or subjective secondary literature, or any other data source that would require further triangulated evidence to verify. As such, we applied no modifier to this type of evidence, on the understanding that evidence from one stakeholder would need to be cross-referenced and validated by evidence from other stakeholders or sources. If no data or argument supported the evidence point or if relevant contrary evidence was provided, the evidence was not coded but still used as information to improve the further analysis.

Table 2 Strength of evidence scoring scheme

Type	Description	Modifier
Verifiable evidence	Refers to data that are both plausible and possible to verify. Such evidence generally describes quantifiable measures that can be physically counted	×2
Plausible evidence	This includes evidence that may make a plausible claim but may draw heavily on assumptions from secondary literature. Alternatively, it may refer to evidence that is the plausible conclusion drawn by an expert stakeholder or observer. Presented evidence may justify this view but lack methodology against which the validity of the conclusion can be verified	×1
Minimal evidence	Some documents may simply claim an outcome but provide no information about the data or methodology used to evidence this claim. Alternatively, a claim may be supported by some evidence, but other contrary evidence is also provided. This evidence was not coded but used to signpost potential data and a need for further analysis	×0

Evidence Saturation

For each sub-statement, we calculated the convergence of all data to score the saturation of its content in relation to the components and how strongly it supported or contradicted the underlying ICMO statement. The convergence was calculated for positive and negative data points using the banding shown in Table 3 to determine

Table 3 Saturation rating

Evidence saturation level	Rating
> 75% convergence of relevant evidence supporting hypothesis	Green saturation level – treated as confirmed evidence
60–75% convergence of evidence supporting hypothesis	Amber saturation level – treated as partially confirmed but level of saturation and divergent views pointed out
< 60% convergence of evidence supporting hypothesis	Red saturation level – not treated as confirmed; discussed within the analysis' findings

Table 4 Categories of ICMO scoring results

Category	Description
Total statement score	Total score for each sub-statement, adjusted for strength of evidence
Total data points	Total number of evidence data points scored against each sub-statement
Average statement score	Average score for each sub-statement, adjusted for strength of evidence
Positive data convergence (%)	Saturation of data in relation to the components and its support of the specific sub-statement
Negative data convergence (%)	Nonsaturation of data in relation to the components and its contradiction to the specific sub-statement
Saturation of data (%)	Predominant data convergence for the specific sub-statement

how different statements would be discussed and analyzed depending on their overall data saturation. Where the majority of evidence for a hypothesis was scored positively, the saturation level would support claims that the hypothesis was accurate; where the evidence was mostly negative, the saturation level supported the opposite, indicating that the hypothesis did not hold true. For the high saturation threshold, more than 75% of the evidence needed to be scored either positive or negative. Low saturation of evidence implied that less than 60% of the evidence scored either positively or negatively.

Coding Results

Table 4 summarizes the score categories that resulted from the coding of evidence against the

ICMO statements in the matrix. Overall, we coded more than 800 individual data points against the ICMO statements, providing almost 2000 total scores. The matrix then generated average scores for each statement or sub-statement, adjusting for the strength of evidence modifiers, which were used to assess the statement's accuracy. The matrix also generated overall saturation scores, which increased confidence in the accuracy of the score achieved. On this basis, the evaluation team formulated the findings of the ICMO analysis.

Table 5 provides three example scores adapted from the evaluation to demonstrate coding results for hypotheses that are shown to be accurate, inaccurate, or divergent due to significant differences in project performances within the portfolio.

Overall, the results of the coding in the CMFP evaluation varied significantly, although this was to be expected. The scatter graph in Fig. 3 shows the average score of all sub-statements coded in relation to their data saturation. High saturation was achieved with around half the sub-statements coded, but very few sub-statements achieved an average score of greater than 1 or less than −1. This is indicative of high reliance on plausible data, often receiving a weaker score with no modifier, unlike verifiable evidence. The chart shows several outliers, highlighting where significantly strong evidence was found supporting the accuracy of the sub-statements.

In the following two charts, the data is split to show specifically the scoring for the interventions (Fig. 4) and the mechanism sub-statements (Fig. 5). For the interventions, all average scores were positive and, with the exception of one outlier, all achieved high saturation and relatively high accuracy scores. We found greater variance and less saturation with the mechanism scores, and on average they received a lower accuracy score with none breaking beyond an average score of 1 or −1. This is not especially surprising as interventions are more observable and verifiable than the mechanisms they hope to trigger, and thus receive higher scores.

These charts say little about the evaluation findings, but they do illustrate several important factors that we explore in the next sections. First,

Table 5 Example scoring results

Category	Hypothesis 1	Hypothesis 2	Hypothesis 3
Total statement score	98	−49	2
Total data points	32	63	44
Average statement score	3.06	−0.78	0.05
Positive data convergence (%)	94	24	54
Negative data convergence (%)	6	76	46
Saturation of data (%)	94	−76	54
Result	Very high likelihood hypothesis is accurate	High likelihood hypothesis is partially inaccurate	Highly divergent evidence indicative of differences at a project level
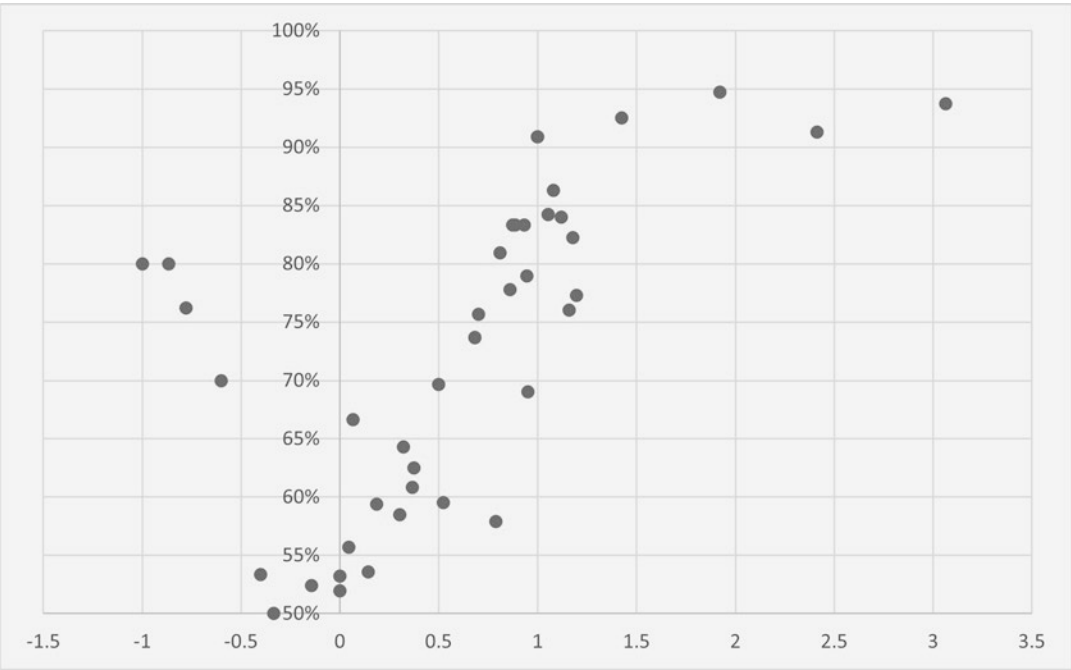
Fig. 3 Average score of ICMO statements in relation to their saturation

more than half the sub-statements received a convergence score of less than 75%, indicating moderate to significant evidence divergence. This was primarily driven by the heterogeneity of the program portfolio, which, due to its small size, could be significantly offset by a single outlying project or small project cluster. Second, few sub-statements scored beyond an average of 1 or −1 for accuracy. This is reflective of the context in which this coding took place, where uncertainty abounds and clear, verifiable evidence is limited. It is also indicative of the maturity of the projects (the latest commencing mid-2018), which limited the availability of strong evidence, especially for mechanism and outcome sub-statements.
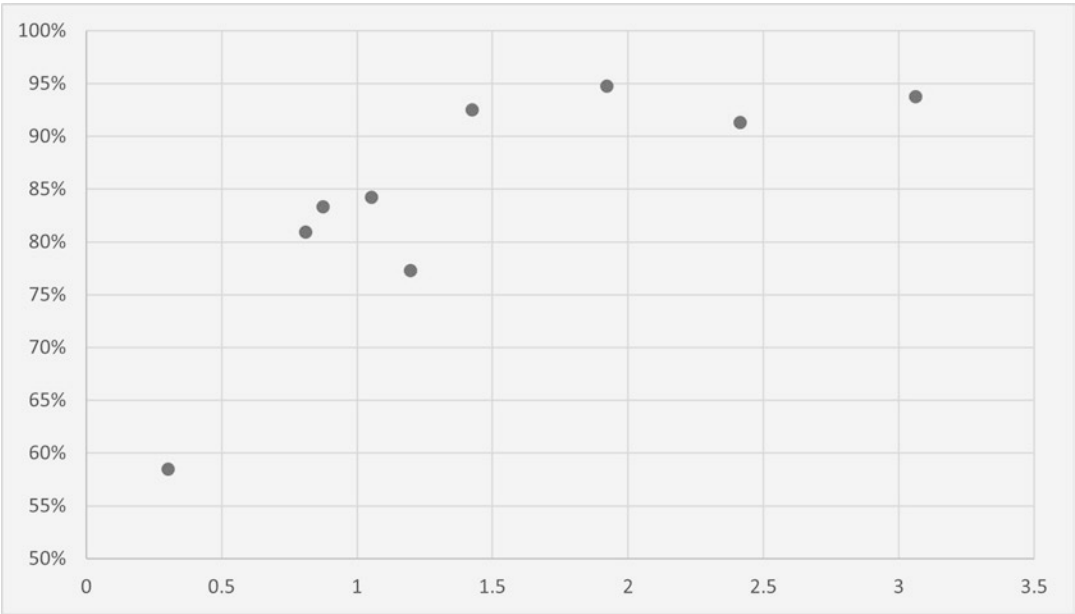
Fig. 4 Average score of sub-statements for interventions in relation to their saturation
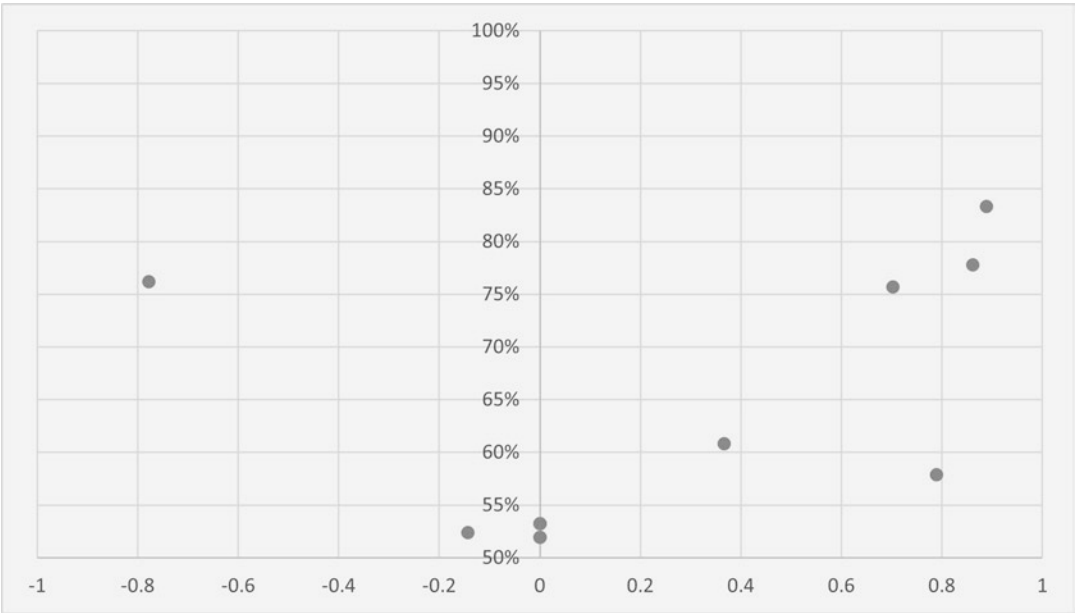

Fig. 5 Average score of sub-statements for mechanisms in relation to their saturation

Realist Evaluation as a Framework

The evaluation team also conducted other descriptive and explanatory analyses to increase the quality of the evaluation findings. To promote consistency, improve the overall robustness of the data scoring, and guide the application of these methods, all data that emerged from these different analyses undertaken were synthesized under the realist evaluation framework and coded accordingly.

To improve the understanding of the program portfolio, we conducted case studies for projects representing half of the portfolio. These covered different technologies, business models, states of energy access markets, and political and regulatory environments. The approach to conducting the case studies was developed using contribution analysis principles, drawing on the six-step guidance developed by Mayne (2008). For the case studies, we collected additional data through interviews with the projects' implementing actors, policy stakeholders, and market experts for each of the technologies represented in the case study portfolio.

Based on the interviews and the additional market information, the team conducted an energy market barrier analysis using Theory of No Change (TONC; Wörlen et al., 2011). TONC is a program-based evaluation approach that looks at the four main groups of stakeholders of the energy access market that can influence the effectiveness of market transformation programs: the users of the technology, the providers of the goods and services (the supply chain), the local and international financiers, and the policy makers. For each of the case study projects, we used a TONC to reveal barriers that impede market change and their intensity. We also performed analysis of how these barriers were addressed by activities of the projects or the program itself.

Finally, to compare how combinations of factors may have contributed to the program's outcomes, the evaluators undertook a qualitative comparative analysis of the program portfolio (Ragin, 2000; Thomas et al., 2014). This is a theory-based approach that applies systematic, logic-based, cross-case analysis to largely qualitative data to identify potential pathways of change (Baptist & Befani, 2015). In particular, it can be used to identify different combinations of conditions necessary to achieve a desired outcome. This is particularly useful in complex settings where contextual and intervention characteristics vary across cases and interdependencies exist between contextual and intervention conditions. The qualitative comparative analysis approach is remarkably compatible with realist thinking—a theory-based approach to

complexity analysis with limited generalizability (Befani et al., 2007)—and provided both a unique avenue by which to analyze evidence regarding causes of project success and evidence generation to parallel and triangulate much of the realist coding (Olsen, 2014).

Benefits of the Applied Approach

Using realist evaluation both as an evaluation method and as the basis for a mixed-methods evaluation framework has several benefits. First, the method offers the possibility of exploring complexity and context in a systematic way. The aim of the evaluation was to analyze program progress and the impact it had on the various local energy markets and the carbon market. Therefore, during the design stage, the evaluation team had to take into account global, national, and local contextual factors and the uncertain outlook of the carbon market. Due to the iterative nature of developing the realist framework, the approach was flexible enough after the initial design stage to be adapted to a changing environment.

During prior evaluation phases, an expansive set of evaluation questions had been formulated. We grouped the questions thematically to address specific areas of interest for this evaluation phase. Based on these groupings, we developed the four ICMO configurations and coded and scored evidence against them. The thematically organized evidence scored under the ICMOs allowed for extraction of findings and recommendations respectively for each of the evaluation questions. This organization also highlighted significant or outlying evidence to provide more nuanced answers to the evaluation questions.

Another benefit is that the configuration of ICMOs is a continuous, retroductive process and is therefore able to consider new insights acquired during the evaluation or major changes of the program's embedded environment. Like the evaluation framework itself, the ICMO configurations can be iteratively adjusted and used for subsequent evaluation phases. For example, if contextual factors have changed or new mecha-

nisms and outcomes are identified, the statements can be modified without requiring a full reset of the evaluation framework. If additional areas of interests arise at a later phase of the evaluation process, evaluators can also develop new ICMO configurations. This ensures that the evaluation methodology is able to keep up with the shifting carbon landscape and developments in its implementation while maintaining rigor and consistency of approach.

A top-down, program-focused approach was used for the ICMO formulation because much time had been spent in the previous evaluation stages to develop and improve the program's theory of change. Developing ICMOs based on program-level information is more time efficient than formulating statements for each project and later abstracting them to make them valid for the whole project portfolio. Program-level statements also make the program theory more visible and give it a clear structure. The abstract structure of the ICMO configurations enabled us to incorporate a wide variety of existing data. To improve the ICMO configurations, the evaluation team refined them based on specific project findings when these findings were also likely to help explain outcomes of other projects. Using the ladder of abstraction, we formulated each specific finding as a general phrase to make it valid for other projects of the portfolio.

The created ICMO matrix allows for high evidence traceability. With the inclusion of the data in the matrix and the developed coding system, we could extract the specific data sources that led to a given finding and categorize the strength of evidence against each finding. We could also filter the evidence regarding individual projects or countries and analyze the ICMO statements separately at the country or project level. This can be advantageous when findings need to be formulated according to different clusters, based on technology, business model or country. For example, in a situation such as the results under Hypothesis 3 in Table 5, the evaluation team could filter data in the matrix to identify which projects were causing the divergence in evidence and investigate those projects further. Once the system is in place, the data is systematically

scored according to its plausibility and importance for the evaluation. In contrast to other evaluation methods, this method minimizes the subjective assessment of individual evaluators: listing the components with signifiers for each sub-statement ensures consistency in the subjective coding process. Moreover, the elaborate framework and coding system can be used for subsequent evaluation phases and can be iteratively improved as the availability of data increases and the understanding of individuals working on the evaluation grows.

Challenges of the Applied Approach

As described above, the realist evaluation methodology has several advantages. However, the development and application of the method during this evaluation also revealed challenges and limitations.

First, the method described requires significant levels of effort. The establishment of the components supporting the accuracy of the sub-statements and the coding process itself were time consuming. Regarding the components, reaching consensus across the team and key stakeholders as to which components would act as signifiers took extensive consultation. The coding process, which is based on a line-by-line isolation approach, required personnel sufficiently familiar with the ICMOs and the evidence base. Had the midterm evaluation not been sufficiently resourced, the approach chosen likely would not have been as effective as it was. In addition to the resource requirements, although the scoring system was highly robust and provided clear, quantitative assessments of hypotheses' accuracy, it was limited in its ability to assess negative findings and did not offer a useful option for assessing which other interventions might have led to a given mechanism, and vice versa.

Second, the coding process was somewhat inflexible, requiring positive or negative scoring against predetermined hypothesis statements, which, if incorrectly set out or phrased, could lead the evaluation team down a narrow path in the wrong direction. Given the highly variable

contexts in which the CMFP operates and the diverse range of outcomes expected at the project level, this approach to coding did not support effective and efficient capture of unstated or unforeseen contextual factors and program outcomes. Although the coding process did include an investigation marker score, the X, to highlight where evidence indicated unpredicted contextual factors or outcomes, in practice this was challenging to implement given the limited opportunity for qualitative description.

Third, the program consisted of 12 projects and developing robust ICMOs for this number proved to be challenging. Working from the bottom up, developing 12 sets of ICMOs—one for each project—would have been time consuming and significantly increased the evaluation data requirements and the ICMO-related consultations. However, the top-down approach, developing ICMOs at the portfolio level and testing them at the project level, was somewhat hindered by the heterogeneity of the projects, which resulted both in ICMO statements that were too general to effectively capture the nuance of the different projects and in high divergence of evidence. This also meant the resulting findings were not always generalizable, and, despite efforts to produce synthesized scores for the wider portfolio, extensive analysis and discussion was required in the evaluation report to explicitly draw out where projects landed on a particular findings curve. Identifying the right balance between specific project-level results versus more abstract portfolio-level results—the right rung on the ladder of abstraction—in such a portfolio is a recognized challenge of realist evaluation and well exampled in this case (Punton et al., 2020).

Finally, although the realist approach addressed many of the challenges created by the uncertain landscape, it could not resolve several fundamental issues. The fact that post-Paris Agreement stakeholder engagement was limited, for example, was not inherently improved by the realist approach beyond highlighting where saturation was low and more evidence needed. The nature of the market also resulted in higher availability of negative evidence, particularly in relation to the transformation of the carbon market, which, if not read and analyzed correctly, could lead to incorrect assumptions about the program results.

Improving the Methodology

Drawing on the lessons learned during the development and application of the realist evaluation approach, this section considers an alternative approach to developing ICMOs and applying realist evaluation in an uncertain landscape with a small but heterogeneous selection of evidence studies.

Bottom-Up Formulation of ICMO Statements

The formulation of the ICMOs themselves may benefit from a bottom-up approach, rather than top-down. Although each of the 12 projects involved had the same basic strategy—using results-based financing and supporting grants to implement commercial business models for energy access technology to generate certified emissions reductions—a core objective of the overall program was to test new business models in different markets, using different technologies. As such, the project models varied substantially, from traditional commercial cookstove sales, to biomass fuel utilities, to public aggregator-led solar home system distribution. Each of these models relies on different intervention strategies, seeks to trigger different mechanisms of change, and must contend with different contextual factors. In cases of such heterogeneity, developing ICMO hypotheses for each project (or cluster of projects, in the case of larger portfolios) is likely to yield more nuanced theories that can be more readily tested at the project level. Had the team developed 12 sets of ICMOs, or even eight sets reflecting the eight project countries, the task at the portfolio level would have been finding the correct level of abstraction at which to synthesize the hypotheses—too high and the aggregated theories risk becoming disconnected to the projects, too low and they will lack generalizability for the wider portfolio. Finding the right balance would allow for the formulation of portfolio-

level ICMOs, which are well suited to assessing the overall success or impact of the portfolio and which can be effectively tested via project studies. However, this approach is not without trade-offs. For example, stakeholder engagement, feedback, and consultation are critical in the effective formulation of ICMOs. Increasing the overall number of ICMOs would increase the engagement requirements, particularly because each project, with a few exceptions, involved an entirely different set of implementors, funders, partners, and other key stakeholders.

Increasing Traceability of Causality by Tailoring the Coding to the Mechanism

ICMO statements are theoretically portable, meaning that a mechanism proven to operate as expected in one situation could feasibly be repeated elsewhere. However, to actually be portable, ICMOs must maintain a balance between having sufficient generalizability to be transferred and enabling appropriate analysis of how and why a given mechanism functioned or was triggered, to ensure any transference or replication is suitably tailored to new contexts (Pawson & Tilley, 1997).

Another option with ICMOs would be to take a different approach to the coding process. In the CMFP evaluation, the team used a coding system that sought to prove the accuracy of each hypothesis and sub-statement. As noted above, this system provided quantitative scoring on whether the overall hypothesis was correct but limited evidence on causality, particularly where the evidence differed from the original theory. In part, this was because each statement type in the ICMOs was treated the same, with equal weighting. An alternative approach would be to place the emphasis in coding on the actual mechanisms of change—the M in ICMO and the critical consideration for the ICMO portability. This approach would still use the concept of signifiers, but would only provide them for the mechanisms. This would reduce the level of effort required to agree on effective evidence thresholds while

allowing for deeper analysis of the causal mechanisms. During the coding process, these signifiers could then be directly tied to the accuracy rating for the whole statement.

In the example in Box 4, the intervention of providing carbon-RBF would ensure "revenue and capacity for projects will be sufficient to overcome the operational challenges in providing rural energy access technologies." Breaking this into signifiers, one might say that evidence of a project-supported company using revenue from the carbon-RBF to recruit new staff or to invest in distribution infrastructure would be good indicators that the mechanism was operating as theorized. Evidence of each of these signifiers would warrant a strong accuracy rating in coding. Thus, the mechanism hypothesis statements could be tested through traditional deductive reasoning.

Box 4: Example ICMO Statement
By providing carbon-results-based financing and business development support funding (I), in a context where there is sufficient customer demand for the energy technologies, access to finance for the pilot enterprises, and a supportive policy framework (C), revenue and capacity for projects will be sufficient to overcome the operational challenges in providing rural energy access technologies (M), resulting in increased energy access and the generation and sale of certified emissions reductions (O).

In this approach, complete ICMO hypotheses should still be constructed, but coding them as a complete unit is not necessary. With the primary focus on providing evidence and proving the mechanism, a more flexible approach applied to the interventions might benefit the overall analysis, particularly given the shifting landscape of the program. Each intervention could be given a tag (I1, I2, I3, etc.) to be used to link the relevant interventions to the mechanism being coded against. With the mechanism example in the previous paragraph, the ICMO was developed with two inter-

ventions: first, the commitment of carbon-RBF (I1); and second, the business development support provided by the program implementor (I2). Using the alternative approach, evidence could be binarily marked positively where it indicates the presence of one or both of these interventions, and scored based on the strength by which the evidence links them to the mechanism. A further consideration is that possible evidence might indicate that the mechanism was due to another intervention (such as the use of the program implementor's influence in the market, an intervention tested by our evaluation team) or an intervention that had not been captured by the original ICMO hypotheses, playing perhaps an even greater role than the originally linked intervention. The original coding approach would not have adequately captured this linkage. In the revised approach, such interventions could be tagged and scored for strength of linkage to the mechanism in question. At the conclusion of the coding process, the evaluation team could then assemble a more complete picture of which interventions contributed to which mechanisms and by how much, based on the actual evidence gathered.

Using this more flexible system allows evaluators to set out their hypotheses at the start of the process and assess how accurate they are, and also allows for effective evidence gathering on causal linkages that had not been drawn at the outset. It supports a more inductive approach to developing causal pathways without the need for the continuous stakeholder consultation that is required for reformulation of the hypotheses.

Increasing Variability of Contextual Factors

In analyzing the contextual factors, a similar inductive approach may be better suited to the uncertainties of the current carbon market. Although certain key contextual factors such as appetite for carbon trading or capacity of implementing organizations were evidently important from the outset, the evaluation team found a variety of surprising and unexpected contextual issues that often proved more critical to the success or failure of each intervention than those identified and coded. Thus, staying open to updating and revising the ICMOs in the face of new evidence that does not fit within the existing framework is important, because narrowing the view of the analysis to specific factors for each mechanism can limit the nuance of the evaluation findings.

Evaluators can employ a revised approach drawing on the evidence-based ICMO assembly method that some favor. The pre-identified contextual factors could be grouped thematically to allow quick coding of the relevant factors demonstrated by specific evidence extracts. This also cuts down on repetition among the contextual factors. In coding evidence, qualitative description linked to a tagged context grouping is likely the most effective option, providing additional detail or analysis on the contextual factors identified by a specific piece of intervention or mechanism evidence. This qualitative coding would also support the capture of other contextual factors not previously identified by the evaluators, thus allowing for both a deductive and inductive approach to coding and developing understanding of the critical contextual factors involved.

A similar approach could be appropriate for outcomes to ensure qualitative capture of unexpected outcomes or outcomes with a stronger link to a given mechanism than envisioned, although generally the bond between mechanism and outcome is more consistent.

Summary of the Modified Methodology

Table 6 presents sample coding that evaluators might apply to the above-described approach. In this approach, only the existence of the mechanism and the links between the mechanism and the interventions or outcomes are quantitatively scored.

The original scoring system would still be applicable to this revised approach for the mechanisms, with each piece of data being scored twice: once for the strength of the evidence (Table 7); and once for the content of the evidence in relation to the signifiers (Table 1). The strength of evidence score applies a multiplier to the content

Table 6 Modified ICMO coding system

Statement	Score	Description
Intervention	Assigned intervention	Yes/no based on intervention established in ICMO
	Strength of linkage	Positive or negative score based on strength of connection to mechanism demonstrated by evidence
	Alternative intervention	Does the evidence indicate that other identified interventions led to mechanism?
	Strength of linkage	Positive or negative score based on strength of connection to mechanism of the alternative intervention
	Other	Does the evidence indicate that non-identified/external interventions contributed to mechanism?
Context	Contextual factors	Which identified contextual factors does the evidence support?
	Other	What other contextual factors are identified by the evidence?
Mechanism	Score	Score the accuracy of the mechanism statement per evidence scoring guide, based on evidence signifiers
Outcome	Assigned outcome	Yes/no based on outcome(s) established by ICMO
	Strength of linkage	Positive or negative score based on strength of connection to the mechanism demonstrated by evidence
	Alternative outcomes	Does the evidence indicate that other identified outcomes are caused by the mechanism?
	Other	Does the evidence indicate that unexpected outcomes are linked to the mechanism?

Table 7 Modified ICMO evidence scoring guide

Score	Definition
3	Evidence strongly supports mechanism statement. Multiple or all signifiers are met.
1	Evidence partially supports mechanism statement. One or more signifiers are met.
−1	Evidence partially contradicts or disproves mechanism statement. Evidence disproves or creates doubt in one or more signifiers.
−3	Evidence strongly contradicts or disproves mechanism statement. Multiple or all signifiers are countered or disproved.

score, recognizing that verifiable and authoritative sources provide more convincing evidence than plausible, subjective sources (see Table 7).

The modified coding system will generate several scores including the overall data score, the total data points, the average data score, and the data saturation in the form of positive and negative data convergence (a score reflecting what percentage of the total data points were positive or negative) and the total evidence convergence. The revised approach to coding also allows for the inductive generation of alternative hypotheses, allowing users to efficiently reflect on the data gathered to reformulate and revise ICMO statements. Further, it provides a reference matrix linking evidence to findings during substantive evaluations.

Conclusion

The application of realist evaluation is appropriate when considering interventions operating in highly unstable or unpredictable landscapes. The method offers a sufficient balance of evaluation rigor and adaptability, allowing for retroductive analysis that can evolve over time. The use of ICMO statements allows for nuanced hypotheses to be tested (provided the right level of abstraction is achieved), which incorporates critical contextual factors at their core. The approach offers the opportunity to find out not only what has happened, but how it happened and, to a lesser extent, why. In the CMFP evaluation, this contextualized understanding was important to generating effective, balanced findings that fairly accounted for the market uncertainty in evaluating the outcomes and impact of the program interventions. The approach developed by the evaluation team for the CMFP was very effective in providing traceable, quantified findings. It also provided a robust framework against which other evaluation exercises could be designed, implemented, and scored.

However, as with all evaluation methodologies, realist evaluation is not without its limitations, some of which were apparent in this evaluation. The significant resource requirements, risks of overly narrow lines of analysis, and challenge in ensuring generalizability are all important lessons that should be considered when setting out to conduct a realist evaluation. Based on these learnings, we have presented a revised approach to conducting realist evaluation. This approach seeks to increase the flexibility of realist evaluation, ensure more nuanced analysis of the causal linkages between interventions and mechanisms, and open the approach further to unforeseen contextual factors or program outcomes.

This approach is unlikely to be suitable for all evaluators seeking to apply realist approaches, nor would it be appropriate for all evaluations. Nevertheless, the following key lessons are useful insights for any evaluator embarking on a realist evaluation:

- **Find the right level of abstraction for your ICMO statements**: When dealing with a portfolio of projects or interventions of any size, setting the right balance between portfolio- and project-level hypotheses is vitally important. Start from the bottom where possible, cluster projects by intervention type if needed, and remember to think about generalizability and portability of the hypotheses.
- **Engage stakeholders regularly but appropriately**: Stakeholder input to the ICMO development process is one of the realist evaluation pillars and consultations should be held at all key development stages, including iterations after data collection. However, striking a balance is important between sufficient engagement and the resource implications, not to mention the evaluator's biggest concern: burdening the commissioner. Agreeing on the process for consultations early in the ICMO development and testing adjustments before consultation may help to strike this balance.
- **Do not underestimate the resource requirements**: Developing, coding, and analyzing ICMOs using either of the methods described above is a time-consuming process. It requires not only individual subjective assessments for every piece of data collected, but also extensive stakeholder consultation and rigor in data gathering. Even if the described coding approach is not adopted, the development of ICMO statements, especially for portfolios of projects, is a difficult task that requires sufficient resources and time for feedback and iterations.
- **Remain open to emerging concepts (and do not be overly deductive):** Developing hypotheses and scoring against them can result in tunnel vision, blinkering evaluators to emerging concepts and data. The retroductive approach described in this chapter, which allows for regular feedback loops between the evidence and the hypotheses, is a beneficial way of thinking about and analyzing the data that allows for multiple iterations of development and the incorporation of emerging ideas and trends.

References

Baptist, C., & Befani, B. (2015). *Qualitative comparative analysis – A rigorous qualitative method for assessing impact*. Coffey.

Befani, B., Ledermann, S., & Sager, F. (2007). Realistic evaluation and QCA: Conceptual parallels and an empirical application. *Evaluation, 13*(7), 171–192. https://doi.org/10.1177/1356389007075222.

Climate Policy Initiative & LTS International. (2018). *Climate public private partnership (CP3) monitoring and evaluation mid-term evaluation*. UK Foreign, Commonwealth & Development Office. http://iati.dfid.gov.uk/iati_documents/50221453.pdf

Department for International Development and Department of Energy & Climate Change. (2013). *International Climate Fund business case and intervention summary: Carbon Market Finance Programme.* https://www.gov.uk/government/publications/international-climate-fund-business-case-and-intervention-summary-carbon-market-finance-programme.

Greenhalgh, T., Pawson, R., Wong, G., Westhorp, G., Greenhalgh, J., Manzano, A., & Jagosh, J. (2017a). *Retroduction in realist evaluation.* RAMESES II Project, National Institute of Health Research, Health Services and Delivery Research Programme. http://www.ramesesproject.org/media/RAMESES_II_Retroduction.pdf

Greenhalgh, T., Pawson, R., Wong, G., Westhorp, G., Greenhalgh, J., Manzano, A., & Jagosh, J. (2017b). *What is a mechanism? What is a programme mechanism?* RAMESES II Project, National Institute of Health Research, Health Services and Delivery Research Programme. http://www.ramesesproject.org/media/RAMESES_II_What_is_a_mechanism.pdf

LTS International. (2020). *Evaluation of the carbon market finance programme*. UK Department for Business, Energy and Industrial Strategy. https://devtracker.fcdo.gov.uk/projects/GB-GOV-13-ICF-0025-CiDev/documents

Mayne, J. (2008). Contribution analysis: An approach to exploring cause and effect. *ILAC Brief 16*, 4. https://hdl.handle.net/10568/70124

Olsen, W. (2014). The usefulness of QCA under realist assumptions. *Sociological Methodology, 44*(1), 101–107. https://doi.org/10.1177/0081175014542080.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Sage.

Punton, M., Vogel, I., Leavy, J., Michaelis, C., & Boydell, E. (2020) *Reality bites: Making realist evaluation useful in the real world*. CDI Practice Paper 22. IDS. https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/15147

Ragin, C. C. (2000). *Fuzzy-set social science*. University of Chicago.

High-Level Panel on the CDM Policy Dialogue. (2012). *Climate change, carbon markets and the CDM: A call to action*. CDM Policy Dialogue. http://www.cdmpolicydialogue.org/report/rpt110912.pdf

Thomas, J., O'Mara-Eves, A., & Brunton, G. (2014). Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: A worked example. *Systematic Reviews, 3*(67). https://doi.org/10.1186/2046-4053-3-67.

United Nations Framework Convention on Climate Change. (2015). Article 6. *Paris Agreement*. https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf

Wörlen, C., Rieseberg, S., & Lorenz, R. (2011). *The theory of no change*. Arepo Consult. https://arepoconsult.com/wp-content/uploads/2020/02/2016_Woerlen-et-al_The-Theory-of-No-Change_International-Energy-Policies-Conference.pdf