

Why Successful Policy Pilots Are (Not) Scaled Up: The Case of European Employment-Creation Schemes

NASTASSIA HARBUZOVA, TALLINN UNIVERSITY OF TECHNOLOGY, RAGNAR NURKSE DEPARTMENT OF INNOVATION AND GOVERNANCE, ESTONIA

Abstract

While existing research demonstrates that successful policy pilots often remain unscaled, the reasons for this observation lack clarity and systematisation. This study addresses the gap. Drawing on the synthesis of various literatures, it first develops four propositions explaining why successful pilots may not scale up in practice, including: success may not be a scaling prerequisite (1), success may not be proven to decision-makers (2), success may lack further funds and incentives (3), and success may not be scalable (4). To “probe” and strengthen the propositions, the study then employs them as lenses to analyse a comparative case-study of three European employment-creation (pilot) schemes: Zero Long-Term Unemployment Territories (France), MAGMA Job Guarantee (Austria), and Kinofelis Ergasia (Greece). Empirical analysis draws on interviews with program stakeholders facilitated with follow-up discussions and desk research. The findings illustrate that: the success of policy pilots is neither a sufficient nor a primary condition for their upscaling (a), and whether the solutions “that work” find further adoption can be largely determined by the interplay of policymakers’ ideological preconceptions and the institutional design of pilot interventions (b).

Keywords: experimental policymaking, policy pilots, policy success, scaling-up, employment-creation schemes

1. Introduction

The growing turbulence of the modern world puts policymakers in a quandary where regular policy solutions no longer work in the face of rapidly unfolding challenges, while their relevant alternatives are yet to be developed. In that context, adopting policy pilots as an approach to designing effective policy responses has gained momentum¹ (Hughes, Yordi and Besco 2020; OECD 2020; Varazzani et al. 2023). Policy pilots exemplify one of the core types of policy experimentation (along with design experiments, experimental legislation, and regulatory sandboxes) and can be defined as time- and scope-limited interventions intended to test

¹ While the opportunities and pitfalls of piloting differ across policy domains due to a number of ethical, legal, political and technical reasons (e.g., Cinar et al. 2019; Liarte et al. 2024; Raudla et al. 2023), a growing use of experimental practices has been observed both in the areas where experimentation have been adopted as a common approach for designing innovations (e.g., healthcare, environmental, urban policies), as well as in those where it previously got only limited application – such as welfare- and employment- policies (Markowitsch and Scharle 2024; Nesti, Cittadini and Bassoli 2025; Varazzani et al. 2023).

novel policy solutions and use evidence obtained to decide their future (Bailey, Hodgson and Checkland 2019; Raudla et al. 2024). In this study, both randomised controlled trials (RCTs) and non-randomised interventions are considered policy pilots as long as they fit that definition.

According to the rational understanding of policy experimentation – which, as Huitema et al. (2018) argue, still dominates the literature – policy-piloting is a process with an attractively linear logic. It implies that if experimental evidence proves a pilot policy successful, the latter will be upscaled – i.e., extended in time, scope, or both (e.g., Stoker 2010). Yet, what the rationale expects pilots to be often runs counter to what the reality shows they are. As a growing body of empirical scholarship shows, many successful policy pilots remain unscaled due to the multiple considerations of political, institutional, administrative, relational, and technical nature (Corduneanu-Huci, Dorsch and Maarek 2021; Mauw, Smith and Torrens 2023; Nair and Howlett 2016). These considerations, however, are scattered across different studies and policy domains, preventing a holistic, empirically informed understanding of both the pragmatic grounds underlying the upscaling decisions and the role that the success of policy pilots plays among these grounds. This study addresses the gap by developing a systematic perspective on the relations between *success* and *scaling-up* in the practice of policy-piloting.

The aim of this study is two-fold: to identify and systematise the reasons that explain a limited record of upscaling effective policy innovations (1), and to strengthen this knowledge further with evidence from the relevant comparative case-study (2). With that, the research question of the study is: why are successful policy pilots scaled up and *are not* scaled up?

To address the question, the study proceeds in two steps. First, it conducts a systematic review of existing literatures and synthesizes four discrepancies between the rationale ("scaling-up in case/because of success") and the practice of policy-piloting, as follows: *success may not be a scaling prerequisite (1)*; *success may not be proven to decision-makers (2)*; *success may lack further funds and incentives (3)*; and *success may not be scalable (4)*. To develop these propositions, the author has examined the existing scholarship, asking the following sub-questions: (a) is the policy pilot success *always* a factor for its upscaling?; (b) if not – *when* is it a factor?; and (c) what are the *other factors*? How, when and why do the other factors affect the prospects of effective pilot solutions? The literature reviewed covers studies on experimental- and evidence-based policymaking, public innovation, and the politics of policymaking. The studies span multiple innovation areas and policy domains, including environmental, health, welfare, and fiscal policies, and administrative, urban and social innovation.

Second, the study examines how each of the four discrepancies manifests itself in a specific policy case – the case of *employment-creation schemes*. The discrepancies serve as analytical lenses to compare the adoption experiences of three (pilot) programs: *Zero Long-Term Unemployment Territories* in France (hereinafter: TZCLD, from *Territoires Zéro Chômeur de Longue Durée*), *MAGMA Job Guarantee* in Austria (hereinafter: MAGMA), and *Kinofelis Ergasia* in Greece (hereinafter: Kinofelis). The program's pilot(ed) is a conceptually similar tool that is commonly labelled as a *job guarantee*, an active labour-market policy aimed at eradicating (long-term) unemployment by creating *voluntary, individually tailored* and *locally beneficial* job opportunities that do not compete with the regular economy and provide at least a living wage (e.g., Tcherneva 2019). The comparative case-study analysis enables us to verify and

strengthen the discrepancies further, which is crucial given their "cumulative" formation. By comparing the theoretical propositions with empirical insights, the study explores the role of *context* in policy-piloting, identifying its potentially key elements for the upscaling trajectories of the employment-creation schemes analysed, both policy (domain)- and program-wise.² As sources of empirical data, the study draws on semi-structured interviews with 18 program stakeholders (2023-2024; primary data source), complemented with desk research and follow-up discussions (with program leadership and evaluators; 2024-2025) intended to verify and enhance intermediate findings.

This study is structured as follows. Section 2 outlines the existing knowledge on the "success-scaling" relations, specifies both terms and introduces the discrepancies developed. Section 3 provides an overview of the programs analysed, and Section 4 outlines the methodology. Empirical analysis is provided in Section 5, and its key findings and inferences are discussed in Section 6. Section 7 concludes the study, indicates its limitations and suggests avenues for further research.

2. Literature Overview: Success and Scaling-up in Policy-Piloting

2.1. Research gap and definitions

Although policy pilots are recognised as primary examples of evidence-based policymaking (e.g., Ko and Shin 2017), the complexity of the real world hardly leaves them a chance to evolve according to the linear scenario where an experimental solution proven successful becomes regular. While a growing body of empirical research supports this claim, the evidence lacks a systematic understanding. Two circumstances account for that.

First, the questions posed in existing studies tend to treat success and scaling-up as two separate outcomes of policy experimentation. Specifically, the literature that (in)directly suggests that not all successful experimental policies are able or intended to be scaled up can be conditionally divided into three overlapping yet self-contained thematic groups. The first group focuses on the issues of policy evaluation, indicating that so far, there is no commonly recognised notion of "success" in experimental policymaking (Checkland et al. 2021; Martin and Sanderson 1999; Stoker and Evans 2016). The second group investigates the challenges in policy experimentation at its different stages – inferring that multiple considerations may prevent policy experiments from evolving according to the rationale (Cairney and Oliver 2017; Reymond, Chandragiri and Ulrich 2020; Vreugdenhil and Rault 2010). The third group of studies explores how policy experiments influence policymaking and what this influence depends on, providing an overview of the factors that affect different outcomes of policy experimentation, including policy scale-up, policy learning, and knowledge diffusion (Baird et al. 2014; McFadgen and Huitema 2017, 2018; Vreugdenhil et al. 2009).

Second, the lack of consensus on the definitions of *success* and *scaling-up* in experimental policymaking prevents their systematisation. In the scholarly discussion, the success of

² I.e., distinguishing between the specifics of the policy case/field (a) and the specifics of the (pilot) program (b); in the article: employment-creation schemes (a) and TZCLD, MAGMA, and Kinofelis (b).

experimental policies has various interpretations. As Checkland et al. (2021, 466) note, "a multifaceted set of frameworks for judging policy success emerges, defining success from multiple perspectives, across time and through varied lenses." While in earlier studies, the success of an experimental solution is associated specifically with its positive evaluation (Haynes, Goldacre and Torgerson 2012; Jowell 2003), more recent research tends to include additional dimensions in the definition. For instance, Ettelt, Mays and Allen (2015a) distinguish between the material and political outcomes of policy experimentation, emphasising the ability of the latter "to catalyse implementation" as an essential ground for judging its success. Whereas Checkland et al. (2021) go further and suggest policy success to be a complex concept that incorporates the dimensions of process, outcomes, and effects of experimentation, in both program and political terms. Alternatively, there are studies that associate the success of policy experiments with the very fact of their scale-up (Binz et al. 2016; Reymond, Chandragiri and Ulrich 2020). Similarly, the scope of the term "scale-up" varies across studies. Scaling up of experimental projects tends to be associated with an increase in their scale along one or several dimensions, such as: the number of enrolled participants, geographical area, institutional spread, time horizon, budget, range of activities and interaction with the related programs (Gillespie 2004; Reymond et al. 2018; Reymond, Chandragiri and Ulrich 2020; Sondal 2024). Furthermore, some studies consider scaling-up a form of policy diffusion along with knowledge dissemination (Vreugdenhil et al. 2009), or as a special case of policy "amplification", along with its stabilisation and speeding-up (Lam et al. 2020).

Reflecting on both circumstances, this study zooms in on the relations between the success and scaling-up of policy pilots, specifying the terms in line with the aim set: to identify why *the rationale* of these relations ("scaling-up in case/because of success") may not work in practice. Specifically, *the success of a policy pilot* is defined as an evidence-based assessment of high policy performance made according to the criteria set by the evaluators. The definition accounts for two points. First, it considers the success to be an outcome of *solely and specifically* the pilot evaluation. Second, it acknowledges success to be based on judgments that reflect the perspective of its evaluators on what and how *should be* assessed. *Policy pilot scale-up* is defined as both the process and outcome of policy extension in time and/or scope, including policy expansion and prolongation, regular adoption and maintenance at scale, yet excluding knowledge diffusion to avoid confusing the term with policy learning.

Drawing on these definitions, the rationale for the "success-scaling" relations looks as follows: policy pilots evaluated as successful are expected to be scaled up, i.e., the fact of their success is expected to be a necessary and sufficient condition for the latter. The next subsection contrasts the rationale with evidence from the scholarship to explore why it may not work in practice.

2.2. Discrepancies between rationale and practice

2.2.1. Success may not be a scaling prerequisite

Existing studies question the very conditionality of pilot policy success for its upscaling, i.e., the extent to which the outcomes of pilot adoption stem from the outcomes of its evaluation.

Literature demonstrates that in practice, policy experiments are launched with multiple motivations and for various purposes (e.g., Ettelt, Mays and Allen 2015b; Torrens and von Wirth 2021), thus challenging the rationale with two alternative scenarios.

Under the first scenario, *scaling up appears to be a predefined decision*. Specifically, whether the experimental solution will be scaled up can be decided before the experimental evidence is evaluated (Nair and Howlett 2016; Sondal, Hellquist and Balfors 2024; Vreugdenhil et al. 2009), without evaluating/collecting it (Bailey, Hodgson and Checkland 2019) or when the evidence has a formal or symbolic character and is used to support rather than question the chosen policy objectives (Ettelt, Mays and Allen 2015a,b; van der Heijden 2014). Three general motivations account for that.

First, policy-piloting may be a tool for gaining political advantage or earning credit from the public (Raudla et al. 2025b). Policymakers may use policy pilots to demonstrate the effectiveness of decisions already made (Zurbruggen and Lago 2019), signal to the electorate their commitment to the "rigour of evidence" (Leino and Åkerman 2021; Sanderson 2002) or provide a "shield of argument" if their policy choices face criticism (Corduneanu-Huci, Dorsch and Maarek 2021). Additionally, policymakers may try to scale up the advocated solution before its evaluation is completed if they doubt their authority to do so in due time – e.g., if the experiment goes beyond an electoral cycle (Bueno 2023; Raudla et al. 2024). Second, a phase-wise adoption of the policy may be *labelled* as a "pilot project" when the intervention is initiated to catalyse the implementation of an already decided solution (Behaghel, Macours and Subervie 2019). Adopting a novel policy in steps may be a way to ensure it would work as intended by the initiators (Ettelt, Mays and Allen 2015a; Nair and Howlett 2016) or to stimulate a shift in public perception and achieve higher social acceptance for the tool (Ansell and Bartenberger 2016). Third, policy-piloting in practice may entail the process of policy prototyping when the emphasis is on evaluating *how* rather than *whether* the solution works (Behaghel, Macours and Subervie 2019; Nair and Howlett 2016). In that case, the pilot is launched to fine-tune the policy before rolling it out at scale (Checkland et al. 2021) or to inform its implementation design (Hughes, Yordi and Besco 2020).

Under the second scenario, *scaling up is not a primary goal of experimentation*. In certain contexts, policy pilots may act as "solutions in themselves," which means that their upscaling in case of success is deemed neither necessary nor attainable. For instance, policy-piloting may be used as a tool for conflict avoidance (Raudla et al. 2025b). Politicians may launch a pilot to delay politically unpalatable reforms beyond their term in office to dodge "electorate retribution" (Bailey et al. 2017). Similarly, piloting may serve as a way to replace unpalatable larger-scale policy reforms with a less painful option (Partners for Health Reform Plus 2004). Alternatively, policy pilots may act as "trailblazers" when the primary purpose of the experimentation is to learn about the tool. In that case, policy learning and policy change appear to be rather disconnected experimental outcomes: the tested solution is in an early phase of development, and its adoption in a given environment is not considered feasible (De Wispelaere, Halmetoja and Pulkka 2023; Ettelt, Mays and Allen 2015b).

2.2.2. Success may not be proven to decision-makers

The rationale of policy-piloting tends to underestimate the fact that policy success is a *value judgement* dependent on what is evaluated, how, and by whom (Checkland et al. 2021; Leong and Howlett 2022; Ruggeri 2025). As empirical studies demonstrate, different policy actors may not share a common view on whether the results of the pilot qualify it as successful (Ettelt, Mays and Allen 2015a; McConnell, Greal and Lea 2020; Parth and Nyby 2022) – drawing attention to the issue of “provability” of success in experimental policymaking.

First, policymakers may question pilot policy success if they doubt the quality of its evaluation (Head 2010; McFadgen and Huitema 2017, 2018). While the “credible evidence” is argued to be crucial for catalysing scaling-up of experimental policies (Spicer et al. 2014), it is unclear *what kind of* evidence qualifies as such. Pilot policies can be evaluated in various ways, yet these ways are unlikely to be seen as equally valid (Abramowicz 2019; McFadgen and Huitema 2018; Patton 2003). The policy experimentation literature indicates that policymakers tend to prioritise one type of experimental evidence/design over another. The priority may be given to RCTs (Bravo-Biosca 2020; Webber and Prouse 2018) or non-randomised interventions (Ettelt, Mays and Allen 2015a; Ko and Shin 2017; Raudla et al. 2023) – depending on what is valued most in a given environment: scientific rigour or hands-on knowledge (Ruggeri 2025). Besides, the reputation of evaluators and their affiliation with the pilot may influence the perceptions of evidence credibility (Jacobson and Azzam 2018; Roseland, Lawrenz and Thao 2015). As the studies on evidence-based policymaking show, policy evaluations conducted by external experts are generally seen as more credible than the internal ones (Jacobson and Azzam 2018; Koon et al. 2020; Vetulani-Cęgiel and Cianciara 2025).

Second, policymakers may have different understandings of what constitutes a “successful” pilot policy and thus judge it on different grounds (Head 2010; Marston and Watts 2003; Ruggeri et al. 2020). The differences may stem from both normative choices and political reasons – referring to the issue of the “right” measurements of policy performance in the first case, and the embeddedness of policymakers’ attitudes towards pilot solutions in their evaluations in the second (Koon et al. 2020; Lewis, McGann and Blomkamp, 2020; Ruggeri et al. 2020; Strassheim and Kettunen 2014). As the evidence-based and experimental policymaking literatures show, policymakers tend to prioritise their “hunches” over scientific conclusions when deciding the fate of policy measures (Bravo-Biosca 2020; Capano and Lippi 2017; Ettelt, Mays and Allen 2015a; Raudla et al. 2023). Accordingly, the success of pilot policies is likely to be judged against certain expectations that policymakers already have about them (Ettelt, Mays and Allen 2015a,b; Maddison 2012; Parth and Nyby 2022). This may seriously complicate proving success to decision-makers who are initially sceptical or not in favour of the pilot tool (Ettelt, Mays and Allen 2015a,b; McFadgen and Huitema 2018; Perkiö 2020). Diverse political interests and party programs compound the case further by giving rise to potentially conflicting discourses on what lessons the experimentation has *actually* delivered (Parth and Nyby 2022; Perkiö 2020).

Both considerations contribute to the fact that in practice, multiple evaluations of policy pilots may co-exist and contradict each other (Cairney 2022; Checkland et al. 2021; Pearce and Raman 2014). In particular, experimental evidence may clash with alternative sources and types of knowledge (Cairney 2022; Wesselink, Colebatch and Pearce 2014), which raises

the issue of evidence prioritisation: the outcomes of *which/whose* evaluation *should be* utilised for judging policy success? Literature on evidence-based policymaking identifies the relations of power as a key factor of knowledge hierarchy, arguing that evidence that matches the interests of the most "empowered" policy actor is likely to be decisive in political decision-making (Marston and Watts 2003; Mattocks 2021). This circumstance may produce the scenario where a pilot proven successful for one actor/agency is promoted as a failure and devalued by another (Leong and Howlett 2022; McConnell 2010; McConnell, Greal and Lea 2020).

2.2.3. Success may lack further funds and incentives

Policy pilots can be initiated by various policy actors and at different governance levels, which may largely affect what funding strategies and sources will be applied to support experimentation and what level of financial autonomy the experimenters will enjoy (Cinar, Trott and Simms 2019; Raudla et al. 2025a). As empirical studies illustrate, the lack of financial autonomy among pilot initiators may become a central barrier to adopting successful policy solutions after experimentation (Bello et al. 2018; Raudla et al. 2025a; Sondal, Hellquist and Balfors 2024). The issue may arise within different funding regimes, due to the limited budgets of experimenting agencies, temporary nature of support when experiments are financed from the targeted (international) funds (e.g., European Social Fund Plus), or the resource dependency "by default" when the funding comes from the budgets of higher-level policymakers or private sponsorship (Ko and Shin 2017; Markowitsch and Scharle 2024; Raudla et al. 2025a). In all these cases, the need for a third-party "buy-in" to scale up the pilots makes their prospects dependent on external political will. Recent research supports this inference with multiple examples of successful policy innovations being unscaled as they failed to find and/or sustain such "political will" (Beisheim et al. 2025; Grin 2020; Nesti, Cittadini and Bassoli 2025; Sondal, Hellquist and Balfors 2024). Two circumstances explain much of the issue.

First, the unwillingness of policymakers to take risks and the availability of safer policy alternatives may leave successful pilot policies temporary. Literature on public policy and innovation describe (higher-level) policymakers as risk-averse actors who generally favour the status quo over policy change, as the latter may jeopardise their political advantage in terms of power or reputation among voters (Bello et al. 2018; Hong and Lee 2018; Liarte, Criado and Alcaide-Muñoz 2024). Associating "less risky" with "more politically palatable," policymakers tend to comply with less-than-optimal yet safer policy alternatives (Callander 2011; Hong and Lee 2018) and may thus be reluctant to replace regular policy measures with experimental solutions (Mauw, Smith and Torrens 2023; Turnheim and Sovacool 2020) or opt for adopting incremental innovations over more radical ones (Grin 2020).

Second, policy pilots may clash with the dynamics of political timeframes. As the respective studies indicate, policy pilots may lose their "appeal" during experimentation – due to the changes in government when the pilots overlap several electoral cycles, changes in voter demands, or various shocks that may divert policymakers from their original plans (Bueno 2023; Leino and Åkerman 2021; Mauw, Smith and Torrens 2023; Raudla et al. 2025b). Altogether, the conflicting timeframes of the piloting and its political environment may ultimately "undo" the adoption of successful policy innovations (Bueno 2023; Raudla et al. 2024).

2.2.4. Success may not be scalable

Finally, the rationale of policy-piloting disregards the fact that successful pilots may have limited *scalability* and thus fail to replicate the same results at scale. According to existing discussions, the following considerations may constrain the scalability of interventions: the "smallness" and research design of policy pilots, the diversity of contexts that the target scale encompasses, and the challenges involved in operating pilot solutions at scale.

To start with, the essential characteristics of policy pilot design, such as their limited scope and timeframe, may compromise the external validity of findings they deliver (Bédécarrats, Guérin and Roubaud 2019; Beets et al. 2020; Werner and Riedl 2019). On the one hand, policy pilots risk not capturing those effects that the tested solutions would generate at a larger scale or over a longer timeline (Bravo-Biosca 2020; Raudla et al. 2023). On the other hand, pilots may fail to capture multiple contextual variations that an experimenting country incorporates – including both the uneven distribution of the issues tackled (e.g., when the pilots target "narrow" social/demographic groups or the issues are intrinsic to certain geographical landscapes (Eboreime et al. 2021; Verkerk and Van Buuren 2017)), and the political and legal fragmentations. These circumstances may introduce a higher generalisability bias in larger piloting countries and RCT-designs that seek "optimal conditions" to ensure internal validity (Beets et al. 2020; Muralidharan and Niehaus 2017).

While the uneven distribution of the tackled issues directly challenges the external validity of pilot evidence, political and legal fragmentations constrain the scalability of the latter, affecting the demand for the pilot solutions or the opportunities for their adoption across the country. The dominance of different parties across local governments or diverging political agendas of local and central authorities may cause political palatability of the solutions to vary and thus block their adoption in areas of low political support (Cairney 2022; Jessoula and Natili 2020; Zhu and Zhao 2021). Similarly, the lack of legal consistency between the experimenting and potential receptive jurisdictions (including the lack or inability to provide legal support for policy experimentation within certain administrative units) may prevent the replication of successful pilot policies (Binz et al. 2016; Eneqvist 2022; Koon et al. 2020; Torrens, Johnstone and Schot 2018).

Finally, the capacities required to effectively operate pilot policies at scale may not be available (Bulthuis et al. 2020; Reymond et al. 2018; Reymond, Chandragiri and Ulrich 2020; Sondal, Hellquist and Balfors 2024; Vreugdenhil and Rault 2010). On the one hand, not all skills and infrastructures underlying the experimental process may be feasible to replicate in the target environment (Howlett and Leong 2022; Lewis, McGann and Blomkamp, 2020; Simmons et al. 2007; Verkerk and Van Buuren 2017). As studies indicate, the success of pilot programs may be inseparable from certain intrinsic and barely scalable elements of the experimental process (Eneqvist 2022; Howlett and Leong 2022). The organisational culture and informal routines of experimenting agencies, individual expertise, leadership charisma and certain technical facilities are some examples (Liarde, Criado and Alcaide-Muñoz 2024; Reymond, Chandragiri and Ulrich 2020; Torrens, Johnstone and Schot 2018). On the other hand, scaling-up may entail an unacceptable increase in costs or complexity of operating pilot policies (Reymond, Chandragiri and Ulrich 2020; Sondal, Hellquist and Balfors 2024). The growing network of policy/institutional actors involved with the policies may require specific

complementary capacities to be developed – to ensure effective coordination between them, establish new collaborations and manage resource provisioning (Cinar, Trott and Simms 2019; Eboreime et al. 2021; Vreugdenhil and Rault 2010). The inability to develop such capacities may prove successful pilots to be "unscalable" as well (Howlett and Leong 2022; Sondal 2024).

Altogether, the four discrepancies introduced describe the process of policy-piloting at its different stages: initiation (1), evaluation (2), and adoption of experimental solutions (3, 4) – providing a plausible systematic overview of how the "success-scaling" relations look in practice. Table 1 summarises the propositions.

Table 1: Why successful policy pilots may not scale up: summary of theoretical propositions

Discrepancies between rationale and practice	Reasons identified
Success may not be a prerequisite	Scaling-up may be predefined → piloting for: - voter validation (symbolic role of evidence) - phase-wise adoption - prototyping: to <i>optimise</i> rather than <i>test</i> novel policies Scaling-up may not be a purpose, or pilots as "solutions in themselves": - political tools for conflict avoidance - "trailblazers" (piloting for learning)
Success may not be proven to decision-makers	Pilot success is a value judgement → decision-makers may: - not consider pilot evidence as credible (<i>normative</i> issues of research design, evaluators, and measurement choices) - be negatively predisposed to pilot policies (<i>political</i> issues of biased evidence, evidence (re/de)valuation and "power-based" prioritisation) ↓ Clash of multiple evaluations, relativity of success/failure
Success may lack funds and incentives	Scaling-up requires (access to) resources → policymakers may lack: a) financial autonomy: insufficient disposable budgets; temporary funding; financing from higher-level policymakers/private sponsorship (↓) b) political incentives to buy-in due to: - status quo, risk-aversion and more palatable policy alternatives - loss of "appeal" due to electoral cycles, changes in voter demands, shocks
Success may not be scalable	Pilots may not yield the same results at scale due to: - limited external validity: "smallness" and research design of interventions - political and legal fragmentation: varying agendas, laws and regulations - constraints of operational design: inability to replicate pilot capacities or develop complementary ones (needed to steer upscaled solutions)

3. Case Selection

To examine how the theoretical propositions manifest themselves in a specific policy case, the author conducted a comparative case-study in the domain of employment(creation) policies. In the literature on experimental- and evidence-based policymaking, the insights from that domain remain very scarce, which is likely due to limited experiences of experimenting with such policy solutions, until recently (Markowitsch and Scharle 2024). This circumstance makes the analysis of employment-creation schemes highly beneficial theory-wise, as it enables the study both (a) to discover the *domain-specific* context of piloting the schemes and (b) to reflect on its importance for the discrepancies to occur. This is particularly relevant given the rise of political interest in the "job guarantees" observed in Europe and beyond over the past few years (Markowitsch and Scharle 2024; Tcherneva and Lalucq 2022).

The case study includes three programs that share the following characteristics. The programs were launched as *policy pilots* (according to official media/reports), are conceptually similar (labelled as "job guarantees"), relevant and comparable in time (ongoing/recent programs). All the programs have been promoted as successful in official media/research/reports (e.g., ILO 2018; Lehner and Kasy 2024; Tcherneva and Lalucq 2022; program websites),³ yet their scaling records vary. With such characteristics, the programs strengthen the study further, enabling the examination of why scaling-up a similar pilot solution may follow different scenarios – and thus reflect on the role of the *program-specific* context for the prospects of policy pilots. The programs analysed are introduced next.

3.1. Zero Long-Term Unemployment Territories (Territoires Zéro Chômeur de Longue Durée), France

TZCLD is an ongoing pilot program extended for a second experimental term (2021–2026). It combats long-term unemployment at a territorial level by creating jobs that match the skills of the unemployed with the unmet needs of their community while facilitating social cohesion. The aim of the program is to make effective the right to employment for all French residents by demonstrating that no one is unemployable, and there is no shortage of work and public funds to do so.

TZCLD was initiated in 2016 by four French NGOs (grassroots) as a 5-year pilot program with the scope of 10 experimental territories (referring to local administrative units of up to 10,000 residents); facilitated by top-level politicians and funded (mainly) from the state (central government budget). In 2021, the program was extended for a second term and subsequently expanded to at least 50 new territories. By 2025, TZCLD will have been piloted at 75 territories employing 3,629 residents (as of 02.2025). The program has inspired piloting similar schemes across Europe, including the initiatives in Belgium and Italy (Markowitsch and Scharle 2024).

The operation of the scheme draws on a novel institutional network created as a part of the program. The Companies for Employment Purposes established based on existing social enterprises are the main implementing bodies, while the program governance is provided by the Local Committees for Employment, chaired by local authorities at the territorial level, and

³ Which does not necessarily entail that the program is successful according to definition adopted in the study (2.1). Whether it is the case is analysed in Section 5.

the TZCLD Association (management/coordination), Experimental Fund (finding allocation to the territories) and the TZCLD Observatory (research/evaluation coordination) at the general level. The scheme is iterative and evolves consensually, drawing on an open, diverse and dynamic network of stakeholders.

3.2. MAGMA Job Guarantee (Modellprojekt Arbeitsplatzgarantie Marienthal), Austria

MAGMA represents "the world's first universal job guarantee experiment" (Lehner and Kasy 2024), a preregistered RCT initiated and funded by the local public employment service (PES), and designed in collaboration with Oxford University researchers (its main pro bono evaluators). The program was aimed at eliminating long-term unemployment and improving the general well-being of the area by offering jobs that match the needs of the unemployed and their community. The purpose of the pilot was to provide scientific evidence on policy performance.

The program was conducted in the town-village Marienthal from 2020 to March 2024, creating 150 unconditional, voluntary and non-competitive jobs for long-term unemployed residents. Employment was created in several forms, including jobs in newly established social enterprises, subsidised jobs in the regular economy, and projects developed by the unemployed. The program was not scaled up and discontinued after the pilot term.

The initiating PES and its regular partner, *it.works*, a non-profit provider of labour market services, implemented MAGMA. To ensure credible evidence, no politicians were involved in the official evaluation and/or operation of the scheme.

3.3. Kinofelis Ergasia, or 'Of Common Benefit' (ΚΟΙΝΩΦΕΛΗΣ ΕΡΓΑΣΙΑ), Greece

Kinofelis is an ongoing public employment program led by the Greek Ministry of Labour (MOL) since 2011. The launch of the pilot was motivated by a sharp rise in unemployment observed in the country due to the global and domestic economic crises. Since then, the program has aimed to reintegrate unemployed residents into the labour market, reduce poverty and social exclusion by creating temporary jobs beneficial for both the unemployed and their municipalities. Limited disposable funds (European Social Fund Plus, MOL budget) and high demand from the population made the program selection-based (the chances for being enrolled varied by age, status, and unemployment term) (ILO 2018).

The scheme was co-designed with the Greek Trade Union Confederation in 2011 and the ILO (next pilot phases, 2016–2017). In 2016, its improved version was introduced and subsequently adopted as a regular policy measure nationwide. Overall, the scaling-up included three steps: from 17 (2011–2013) to 51 (2013–2015) and to all 326 municipalities (2016) with 25,000–45,000 participants annually. In 2019, the program was downscaled and has since been conducted in a reduced form.

The implementation of Kinofelis draws on the collaboration of the MOL (defines enrolment

terms, reimburses employment costs) with the municipalities (creates and steers employment projects).

4. Methods

To collect data, the study combines interviews with program stakeholders and complementary desk research. Altogether, 18 semi-structured interviews were conducted (5 for TZCLD, 7 for MAGMA, 6 for Kinofelis; 2023–2024). To select the interviewees, a criterion-based sampling (different roles in the pilot) was mixed with the snowball technique. The sample includes initiators, managers, evaluators and direct implementers of the programs (Appendix 1).

The interview questionnaire was iteratively adjusted based on the interviewees' experience and previously collected data. The questions asked dealt with piloting challenges, details of the evaluation process and results, scaling-up factors and expected/fulfilled scenarios for program adoption, and personal lessons learned by the interviewees during experimentation. The goal of each interview/discussion was to collect specific data, verify intermediate inferences or compare the perspectives of different stakeholders. The interviews of an hour on average, were conducted online, recorded and transcribed. To preserve anonymity, the interviewees are identified by codes in the analysis (e.g., "Fr1/Au1/Gr1"; Appendix 1).

Data analysis followed an iterative and interpretive approach consistent with a small-N case-study design (Creswell and Poth 2023). It was conducted in parallel with data collection so that one informed the other. Specifically, each of the four discrepancies (Subsection 2.2) was alternately "probed" on each of the three programs (Section 3) to analyse *to what extent* the discrepancy resonates with the program and *why*, i.e., whether the factors explaining its occurrence in the case are similar to explanations offered by the literature. In the next section, each case follows the logic described.

In addition, the analysis has benefited from feedback received during follow-up discussions with program stakeholders (program leadership and/or evaluators; 2024–2025). The discussions were held in connection with the author's presentations of intermediate (2024) and general (2025) results of the study on piloting "job guarantees" and enabled the verification and enhancement of the inferences made.

To help the reader follow the analysis (Section 5), Appendix 2 provides the charts that visualise program timelines and key events that might have affected them.

5. Analysis

5.1. TZCLD

The program proved to be the only case where the results of experimental evaluation directly affected its chances for scaling up. The primary motive for launching TZCLD was to demonstrate that the scheme works and ultimately achieve its legal adoption nationwide (Fr1–2, Fr5). While

the initiators themselves expected the scheme to be successful (Fr3-4), scaling up required proving that to the state (central government, MOL) (Fr1). As an interviewee elaborated, "because government is funding a significant part of the project, it is the government's decision: to move it forward, stop it, or reduce it" (Fr2).

The pilot featured two simultaneous evaluations. On the one hand, the scheme was evaluated "from within," by stakeholders. The initiators adopted a broad approach to evidence collection, gathering multiple types of learnings (scientific, hands-on, grassroots-feedback) on the program's effects, outcomes and costs (Fr1, Fr4). Although the evidence indicated the success of the program (Fr1-5) as well as encouraged its scalability, elucidating how the scheme should be iterated to operate across diverse French territories (Fr1), the central government did not consider it credible due to the risk of "dependent" evaluation (Fr1-3). As an interviewee commented, "we [stakeholders] might try to find what we are convinced of" (Fr3). On the other hand, and relatedly, the evaluation was conducted by the Scientific Committee assigned to the project by law (Fr1). This evaluation, in contrast, reflected the priorities of the state, focusing on collecting scientific evidence on the project's cost-benefit ratio (Fr2-3). At the time of the analysis, the outcomes of this evaluation were uncertain, yet the stakeholders were concerned that it might show the program to be a failure due to the high costs (Fr2-4).

Thus, the program faced the challenge of potentially conflicting evidence when the pilot success "according to stakeholders" might be devalued by "external" decision-makers. Furthermore, the interviewees questioned the neutrality of the Scientific Committee (Fr3-4). As one of them noted, the government had room for "tweaking" its conclusions by choosing what the Committee should and should not include in the program cost-benefit ratio: "sometimes we had the feeling that one of the tasks of the Scientific Committee was to kill us – to show how expensive the project could be – while we don't see how externalities are examined" (Fr4). The interviewees linked the lack of state support to the risks that the program posed to the status quo of the MOL. The adoption of the scheme at scale would imply an increase in the autonomy of local-level authorities at the expense of the MOL, including the cuts in disposable funds and reduced control over the territories (Fr4).

While the lack of political support from decision-makers threatened the program's upscaling beyond its legally protected pilot-term (Fr1), the political diversity of French territories complicated expanding the pilot. As interviewees clarified, expanding the experimentation to new territories required reconciling the conflicting interests of different political parties in power in different territories (Fr4, follow-up discussion). To make the pilot ideologically palatable for both "sides," the importance of diversifying political narratives to promote the scheme and adopting a consensual decision-making model for pilot governance was emphasised. Ultimately, such adaptations turned the challenge into a strategic advantage for the project. By demonstrating that "they [left and right] can make the project together" (Fr4), TZCLD neutralised its initially leftish affiliation and grew resilient to changes in the political agenda due to legislative, presidential and European elections (follow-up discussions).

Beyond political concerns, the issues of scalability made the interviewees question the nationwide rollout of the program. The fact that the governance (and in part, implementation) of TZCLD relied on the novel institutions created within the project caused its effective expansion to be particularly difficult. First, scaling up the scheme required defining how to

integrate the novel entities into the system of existing institutions and harmonise their mutually informed work at scale (Fr2). Second, different operationalisations of the scheme across French territories cast doubt on whether the program can be kept transparent nationwide. An interviewee explained, "the challenge will be to ensure that all territories still comply with the key fundamentals of the project: not compete with the normal economy, hiring people without selection. To monitor this at a larger scale will be much more difficult" (Fr2). Finally, the uncertainty of functional overlaps between the program and the recently reformed PES-system compounded the case. The stakeholders had to envision "how it is going to match" (Fr1) and "prove that the experimentation can work with the reform" (Fr4). As a result, a *truncated scaling-up of the program* became a likely scenario (Fr4).

5.2. MAGMA

For MAGMA, scaling up in case of success was neither a prescribed nor an anticipated scenario. The program was launched as a temporary project with an intention to produce rigorous scientific evidence on policy performance that could be used at some point in the future, under a conducive political agenda (Au1-7). With that, MAGMA resembled the case of a "trailblazer" where policy learning and policy change are spaced out in time (2.2.1). Yet in this context, the initiator's limited decision-making autonomy (the authority and funds of local PES sufficed to pilot but not to upscale labour market innovations), coupled with the program's low political palatability, explained the case. As the agency representative noted, using pilots as "solutions in themselves" was a common practice for the agency: "we had different pilots for many years and, to be honest, it always comes back to money. There were really successful projects. But we just didn't have the money to scale them up, so they ended up temporary" (Au4).

The conservative government in office (Au1, Au3-4, Au7) and low unemployment rate in the region (Au5-6) reinforced the expectation that the program would remain unscaled: "the job guarantee program is clearly a social democratic idea. And now the government party [conservatives] is not interested in financing such projects" (Au3). Consequently, it was not a surprise for the stakeholders that evidence of policy success, despite its rigour, would fail to convince the government (Au1, Au3-4, Au7). As an interviewee explained, "even with this scientific proof, they [government] can always say: 'you just measure what you want to measure.' And I don't think they will be very much influenced by it" (Au4). Ultimately, the program was criticised by conservatives on alternative grounds (program financial efficiency) (Au1, Au3-4). MAGMA was declared "costly" (Au4), which could be partly misleading: "they [conservatives] declared to the Vice-Chancellor that the researchers wasted a lot of money, although there was no money involved" (Au1).

Along with the conducive agenda, the interviewees emphasised the importance of political facilitation to resume the program in the future (Au1-2, Au4, Au6-7; follow-up discussion). The advocacy leverage for initially politically neutral pilot was developed "from within" when the (former) head of initiating PES became a leading regional politician (head of the Social Democratic Party). As an interviewee commented, "at some point, you always have to involve politics to push it forward. And now we are at that point" (Au7) [the interview was conducted before the 2024 Austrian legislative election]. The change in leadership, however, created the

risk that the efforts to resume the program might not be supported by the new PES manager, who "is not from social democrats" (Au7).

Importantly, when discussing the program resumption, the interviewees referred to piloting renewal rather than a permanent adoption of the scheme. The initial scope of the pilot (150 residents), affordable within the disposable budget, was cited as a potential constraint on the program's scalability (Au3-6). As an interviewee noted, a better contextual awareness should be developed first: "I'm sure you would first have to adapt it in a city because it's just not the same. To get more learning there and then see if this is the way it should work" (Au4).

5.3. Kinofelis

For Kinofelis, scaling up of the scheme was driven by political commitment rather than experimental evidence (Gr1, Gr4, Gr6). The program's initiator (MOL) had sufficient decision-making autonomy to proceed with adopting it nationwide before the pilot evaluation was completed: "the program was for eight months, but four months later we rolled it out for the rest of the country," commented an interviewee (Gr1).

The ministerial office expected the program to be successful based on findings from prior research (Gr1, Gr3), and there was no need to corroborate these expectations further. Consequently, the evaluation of the program itself was aimed at *optimising* rather than testing the scheme. As several interviewees noted, despite being labelled a "pilot" (e.g., ILO 2018), the experimentation was more akin to a phase-wise adoption of the policy (Gr1) or policy prototyping (Gr2, Gr4), with the evaluation focused on *how* rather than *whether* the scheme works. One interviewee explained that "the aim of evaluation was to check if the project is being implemented right and whether any adjustment needs to be done" (Gr2). Accordingly, the pre-established nature of the program scaling-up left no room for valid criticism of its evaluation or effectiveness during the pilot phases.

The "shield" of political commitment, however, proved to be a short-term privilege. After the leadership turnover (2018) followed by the change in government from the radical left to the centre-right (2019), Kinofelis lost its political appeal (Gr1, Gr4-6). The new ministerial office did not support employment creation, neither ideologically (Gr1) nor did it regard it as an expedient expense item (Gr4). Consequently, the program was *scaled down* in scope, frequency, and quality of enrolments. As interviewees clarified, "they [new ministerial office] decided they would keep it, but very small: like, 30,000 people every 1.5 year" (Gr1), while "the training, the capacity building, the updates to the platforms that we did – they didn't happen in future implementations" (Gr4).

Nevertheless, the proactive leadership during the "pilot-phases" enabled ensuring that the program is (a) scalable (by targeted development of municipal capacities), (b) properly legitimised among regular PES and (c) creates a sustainably high demand from the unemployed (Gr1-2, Gr4-6). As a result, Kinofelis became "so popular among the local population" (Gr1) and "so well-entrenched in the system" (Gr4) that the scenario of its termination proved to be too politically expensive to implement (Gr1, Gr4-5). As an interviewee elaborated, "the new government explicitly stated that they want to abolish Kinofelis, the agenda and priorities have

changed... but they still keep it running, because abandoning it would entail such political costs that it would never be" (Gr4).

6. Discussion

The analysis has illustrated that the relations between the success and scaling-up in policy-piloting tend to be complex, dynamic, deeply contextualised and highly political – reflecting the rationale ("scaling-up in case/because of success") only to a limited extent. To structure and further discuss this empirical observation, the section applies the lenses of four discrepancies developed. Table 2 summarises the discussion.

The first discrepancy – *success may not be a prerequisite for upscaling* – finds support in the cases of Kinofelis and MAGMA. The programs corroborate both major scenarios: a pre-planned adoption and an expected termination of policy pilots. The Kinofelis case demonstrates that scaling up of pilot programs can be predefined when the initiating agency has sufficient decision-making autonomy (authority, funds) for that. Although labelled a "pilot" (ILO 2018), the program's national roll-out was decided already at the stage of its inception. The "pilot phases" and related evaluation were used to *optimise* the novel scheme rather than *test* it. The case of MAGMA, in turn, indicates that policy pilots can be launched as "solutions in themselves" when the initiators anticipate them to remain temporary irrespective of evaluation results. Specifically, the limited budget of local PES and an unfavourable political agenda (right-wing government, low unemployment rate in the region) created the initial expectations that MAGMA would likely be terminated after the pilot.

Along with that, the analysis has shown that policy pilots can pursue multiple purposes at once (as aligns with Ettelt, Mays and Allen 2015b). While the intention "to test" can be among these purposes (as in the case of MAGMA and arguably – TZCLD), pilots can also serve as *policy measures* aimed at resolving (a part of) the issue during experimentation (all three cases). Along with that, pilots can serve as *political tools* utilised to facilitate scaling-up of novel programs or "anchor" them at scale by creating political incentives/costs for the program adoption/termination. In particular, piloting can be employed to build an international support coalition for the novel program (TZCLD), acquire credible evidence to advocate for the program adoption under a conducive political agenda (MAGMA), or use the pilot term to ensure the program is effectively enshrined among regular public services and meets the needs of the target population (to create *sustainable* demand for it).

The second discrepancy – *success may not be proven to decision-makers* – is vividly corroborated by the examples of TZCLD and MAGMA, the programs where the success of the scheme *is to be* proven to higher-level policymakers to proceed with its further adoption. The programs indicate that the normative and political issues of pilot evaluations are deeply intertwined. The lack of common guidelines on *which* aspects of pilot policies should be evaluated can, and is likely to, be used in political interests. In both cases, policymakers had room to influence the outcomes of pilot evaluation irrespective of its rigour – either directly affecting the criteria of the official evaluation (TZCLD) or while criticising the pilot with alternative evidence (MAGMA). Consequently, the analysis suggests that the issue of *power-based* evidence prioritisation may become unresolvable for scaling-up ideologically/

ideationally contested pilots with the limited decision-making autonomy (as aligns with Mattocks 2021; Nesti, Cittadini and Bassoli 2025). Successful according to internal evaluations, neither TZCLD nor MAGMA managed to convince higher-level policymakers, who prioritised alternative evidence that resonated with their scepticism of the programs (respectively: official evaluation regulated by the state (TZCLD) and program financial assessments promoted by conservatives (MAGMA)).

Regarding the third discrepancy – *success may lack further funds and incentives* – the case study *altogether* demonstrates that the financial autonomy of policy pilots is a crucial starting point to understand why their scaling scenarios vary. While in the Kinofelis case, the ideological commitment of its leader *per se* provided sufficient grounds to proceed with scaling up the program nationwide, the prospects of TZCLD and MAGMA (led by no less committed leadership who, however, had no [TZCLD] or limited [MAGMA] financial autonomy) were threatened by the lack of political support from the higher-level policymakers regulating access to the corresponding funds. For TZCLD, the lack of support from the central government (mainly the MOL) complicated expanding the pilot to the new experimental territories and made its stakeholders doubt the chances of the pilot “to survive” after the second experimental term (Fr4). The risk that the program posed to the status quo of the MOL (due to the reallocation of resources from central- to local-level government, which it involved) proved to be a primary reason for that. For MAGMA, the issue of (low) ideological palatability came to the fore. Specifically, the lack of support for “clearly social democratic idea” (Au3) from the MOL and the state-level PES, who followed the agenda of the conservative government, prevented the program from acquiring funds for its extension and/or expansion.

At the same time, the analysis emphasises that political/ideological palatability of policy pilots is dynamic and changes along with the government/electoral cycles (all programs) and leadership turnovers (Kinofelis, MAGMA). For MAGMA, the potential victory of social democrats at the legislative elections was expected to provide a favourable agenda to resume the program, while the change in agency leadership was perceived as a risk for this scenario. For TZCLD, the various-level elections (legislative/presidential/European) were anticipated to influence the “appeal” of the scheme for the state, shaping its support for the program accordingly. For Kinofelis, the change in government from left to right and leadership turnover caused a substantial downscaling of the program. Yet, the case has also shown that the policymakers’ reluctance to make electorally inexpedient decisions may allow for preserving politically/ideologically unpalatable (pilot) programs – suggesting that the tendency of policymakers to favour the status quo over policy change (2.2.3) may *also*⁴ work “for” the pilot upscaling.

The fourth discrepancy – *success may not be scalable* – is supported by MAGMA and TZCLD. For MAGMA, the small scope of the pilot and, in part, its dominant approach to evaluation (preregistered RCT) questioned the feasibility of the program’s national roll-out at the next step. While discussing the desired scenario for the pilot under a conducive political agenda,⁵ the interviewees referred to promoting the experimentation to other Austrian territories, since “a better awareness of local contexts should be developed first” (Au4). Besides, the case also indicates that the research design of policy pilots may stem from budgetary constraints

4 In contrast to literature insights in 2.2.3.

5 Where social democrats would have won the most votes in the 2024 Austrian legislative election.

rather than (solely) normative preferences of initiating agencies, implying that the limited (access to) funding may constrain pilot scalability as well.

TZCLD corroborates the discrepancy in several respects. First, the high political polarisation of policymakers at the local level (across and within French territories) challenged the scalability of the program. The expansion of the pilot with initially "leftish" connotation risked not happening, had its leaders failed to appeal, reconcile and ensure effective collaboration of policymakers on both sides of the political spectrum. The case, however, also indicates that the challenge of political fragmentation/polarisation is resolvable and, moreover, could provide strategic opportunities for the prospects of the pilot programs (such as the resilience to political cycles, "demonstration effect" facilitating pilot expansion, stronger advocacy leverage to request resources from the state). Second, the program shows that the institutional set-up of policy pilots may constrain their scalability. Unlike MAGMA and Kinofelis, the implementation and governance of the French employment-creation scheme relied on novel institutions that first had to be set up in each experimental territory. As a result, the stakeholders were worried it might not be feasible to implement the scheme in the same form nationwide due to the issues of integration (novel vs existing institutions), coordination and transparency.

Table 2: Why successful policy pilots are not scaled up: empirical summary

Discrepancies between rationale and practice	Reasons identified
Success may not be a prerequisite	Scaling-up may be predefined: piloting for prototyping /adopting phase-wise (Kinofelis) Scaling-up may not be <i>the only</i> purpose /pilots as "solutions in themselves": - policy measures: piloting for solving issues during experimentation (all) - political tools/facilitators: piloting to create incentives/termination cost (all) - "trailblazers" (MAGMA)
Success may not be proven to decision-makers	Relativity of success and failure in policy-piloting (TZCLD, MAGMA): - normative and political issues are intertwined: the lack of evaluation guidelines can be used in political interests - ideological/ideational preconceptions are embedded in evaluations - issue of multiple evidence: conflicting evidence can coexist - issue of evidence prioritisation: pilot evidence can be devalued with alternative, more palatable
Success may lack funds and incentives	Limited financial autonomy of pilot initiators against low political support (TZCLD, MAGMA): - limited disposable budget and ideological incompatibility with the ongoing government agenda (MAGMA) - dependence on state funding and threat posed to the status quo (power, disposable resources) of decision-makers (TZCLD) Loss of appeal due to government/election cycles and leadership change (all)

Success may not be scalable

- Pilots may not yield the same results at scale:
- limited external validity: scope, evaluation choices, budgetary constraints (MAGMA)
 - political fragmentation/polarisation: varying "leanings" of local policymakers/ local-level dominance of different parties (TZCLD)
 - constraints of operational design: reliance on experimental institutions and related challenges (integration, coordination, transparency at scale) (TZCLD)

7. Conclusions

This study has systematically explored why successful policy pilots *are* and *are not* scaled up in policymaking realities. Drawing on the synthesis of existing scholarship, the study developed four theoretical propositions for the matter and applied them as lenses for subsequent case-study analysis. The analysis corroborated all the propositions, illustrating that the rationale of "scaling-up in case/because of success" in practice is subject to numerous limitations. On the one hand, pilot evidence that the policy "works" may not be necessary, convincing or politically appealing to proceed with its adoption beyond experimentation. On the other hand, successful as a pilot, the policy may fail to replicate the same results at scale.

While all four propositions found support in the comparative case-study, the reasons behind them (Table 2) only partially overlapped with the ones suggested by the literature (Table 1), were prioritised in a certain way and/or varied among the programs studied – indicating the importance of the piloting context for the prospects of novel policies. By comparing theoretical propositions with empirical findings (Section 6), the study has identified the key *policy*- and *program*-specific contextual factors which, while interacting with one another, prove to shape scaling-up of three employment-creation schemes analysed.

As a key policy-specific factor identified, the study emphasises the ideological/ideational preferences of policymakers. In all three cases, policymakers' preconceptions about employment-creation schemes motivated the launch of the pilots (referring to the initiators' commitment to the schemes), affected their evaluation results (referring to conflicting evidence (interpretation) and its embedded bias) and ultimately, largely influenced the programs' scaling-up – facilitating or hindering their extension, nationwide adoption and maintenance at scale. By contrast, the importance of ideology for the outcomes of policy experiments has received only limited attention in the existing discussions.⁶ This is likely linked to a higher political complexity and thus the lack of experimenting with those policy solutions or in those policy domains/areas which involve high polarisation of political opinions over the issues tackled or approaches utilised for that – as the case-study exemplifies and aligns with Bécharde et al. (2023).

As a key program-specific factor, the study indicates the importance of the institutional design of policy pilots. Specifically, such pilot characteristics as their modes of launch (top-down vs bottom-up) and related decision-making/financial autonomy of initiators, types of *institutions employed to operate them* (conventional vs experimental entities, established

⁶ With few identified exceptions examining the role of party leanings and/or narrative diversification in promoting (experimentation with) minimum/basic income schemes (Jessoula and Natili.2020; Parth and Nyby 2022; Van Parijs and Vanderborght.2017).

vs new collaborations), and *learning strategies adopted* (including the choices of research design, evaluation focus and evaluators) proved to shape their scaling-up scenarios – as resonates with the findings of McFadgen and Huitema (2017, 2018).

Altogether, the differences identified between (a) the programs analysed and (b) the theoretical propositions and empirical findings in no way claim to be attributed *solely* to the context of policy pilots, but may also stem from the limitations of empirical data (such as its interpretative nature, inability to "probe" all the reasons that could underlie the discrepancies, lack of access to certain interviewees), limits of literature review or knowledge gaps in existing scholarship. Besides, the overall generalisability of the findings offered is likely to be limited given the inductive nature of the study and the research design applied (small-N comparative case-study, N = 3).

Thus, future studies would play an important role in strengthening the findings of this study further – in line with the continuously evolving knowledge formation in qualitative research. The framework of four discrepancies could be employed in future comparative investigations to continue holistic exploration of the factors, mechanisms and patterns of scaling-up (or lack thereof) in experimental policymaking. Besides, more research is needed to explore how the choices of institutional design affect the prospects of policy pilots, and how politics mediates the way pilot policies are institutionalised.

In future research on scaling-up of experimental policies, the study proposes to conceptualise "scaling-up" in broader terms, i.e., as a *lasting two-dimensional process*.⁷ Taking a longer perspective would enable to explore the factors that may: (a) have different effects depending on the timeframe considered (e.g., risk-aversion of policymakers, political diversity of pilot stakeholders), and (b) cause reduction or termination of pilot policies after their adoption at scale – embracing the issues of "downscaling" as another essential part of the pilot adoption journey – as aligns with Sondal, Hellquist and Balfors (2024). Incorporating both time-and scope-dimensions in the analysis would allow distinguishing between the expansion and prolongation challenges, which, although largely overlapped, could be different in their substance, effects or reasons.

Acknowledgments

This study was supported by the Estonian Research Council grant PRG1125.

References

- Abramowicz, M. 2019. "Tax Experimentation." *Florida Law Review* 71(1), 65–122.
- Ansell, C. and M. Bartenberger. 2016. "Varieties of Experimentalism." *Ecological Economics* 130, 64–73.
- Bailey, S. et al. 2017. "The Policy Work of Piloting: Mobilising and Managing Conflict and Ambiguity in the English NHS." *Social Science & Medicine* 179, 210–217.

⁷ Contrary to existing scholarship that mainly associates the term with the *piloting expansion* (e.g., adopting similar pilots to other municipalities) or the *ultimate outcome* of experimentation (such as permanent adoption of the novel tool) – see Reymond et al. 2020; Spicer et al. 2014; Vreugdenhil and Rault 2010 as examples for both.

- Bailey, S., D. Hodgson and K. Checkland. 2019. "Pilots as Projects: Policymaking in a State of Exception." In D. Hodgson et al. (eds). *The Projectification of the Public Sector*. UK: Routledge, 130-148.
- Baird, J. et al. 2014. "Learning Effects of Interactive Decision-Making Processes for Climate Change Adaptation." *Global Environmental Change* 27, 51-63.
- Bédécarrats, F., I. Guérin and F. Roubaud. 2019. "All that Glitters is not Gold. The Political Economy of Randomized Evaluations in Development." *Development and Change, International Institute of Social Studies* 50(3), 735-762.
- Beets, M. et al. 2020. "Identification and Evaluation of Risk of Generalizability Biases in Pilot Versus Efficacy/Effectiveness Trials: A Systematic Review and Meta-analysis." *International Journal of Behavioral Nutrition and Physical Activity* 17, 19.
- Behaghel, L., K. Macours and J. Subervie. 2019. "How can Randomised Controlled Trials Help Improve the Design of the Common Agricultural Policy?" *European Review of Agricultural Economics* 46(3), 473-493.
- Beisheim, M. et al. 2025. "Politics Matters! Political will as a Critical Condition for Implementing the Sustainable Development Goals." *Earth System Governance* 24: 100244.
- Bello, B. et al. 2018. "Does Austerity Drive Public Service Innovation? Evidence from Shared Senior Management Teams in Local Government." *Public Money & Management* 38(2), 131-138.
- Béchar, B. et al. 2023. "Political Complexity and the Pervading Role of Ideology in Policy-Making." *Journal of Dynamic Decision Making*, 9(1).
- Binz, Ch. et al. 2016. "The Thorny Road to Technology Legitimation. Institutional Work for Potable Water Reuse in California." *Technological Forecasting and Social Change* 103, 249-263.
- Bravo-Biosca, A. 2020. "Experimental Innovation Policy." *Innovation Policy and the Economy* 20, 191-232.
- Bueno, N. S. 2023. "The Timing of Public Policies: Political Budget Cycles and Credit Claiming." *American Journal of Political Science* 67(4), 996-1011.
- Bulthuis, S. et al. 2020. "Factors Influencing the Scale-up of Public Health Interventions in Low- and Middle-income Countries: A Qualitative Systematic Literature Review." *Health Policy Plan* 35(2), 219-234.
- Cairney, P. 2022. "The Myth of 'Evidence-Based Policymaking' in a Decentred State." *Public Policy and Administration* 37(1), 46-66.
- Cairney, P. and K. Oliver. 2017. "Evidence-based Policymaking is not like Evidence-Based Medicine, so how far should you go to Bridge the Divide Between Evidence and Policy?" *Health Research Policy and Systems* 15, 35.
- Callander, S. 2011. "Searching for Good Policies." *American Political Science Review* 105(4), 643-662.
- Capano, G. and A. Lippi. 2017. "How Policy Instruments are Chosen: Patterns of Decision Makers' Choices." *Policy Sciences* 50(2), 269-293.
- Checkland, K. et al. 2021. "'Success' in Policy Piloting: Process, Programs, and Politics." *Public Administration* 101(2), 463-480.
- Cinar, E., P. Trott and C. Simms. 2019. "A Systematic Review of Barriers to Public Sector Innovation Process." *Public Management Review* 21(2), 264-290.
- Corduneanu-Huci, C., M. Dorsch and P. Maarek. 2021. "The Politics of Experimentation: Political Competition and Randomized Controlled Trials." *Journal of Comparative Economics* 49(1), 1-21.
- Creswell, J.W. and C. N. Poth 2023. "Qualitative Inquiry and Research Design: Choosing Among Five Approaches." Thousand Oaks, CA: SAGE. 552.
- De Wispelaere, J., A. Halmetoja and V.-V. Pulkka. 2023. "A Primer on the Finnish Basic Income Experiment: From Design and Implementation to Evaluation and Impact." In M. Torry (ed.). *The Palgrave International Handbook of Basic Income*. Cham: Palgrave Macmillan, 413-435.
- Eboreime E. A. et al. 2021. "Appraising and Addressing Design and Implementation Failure in Global Health: A Pragmatic Framework." *Global Public Health* 16(7), 1122-1130.

- Eneqvist, E. 2022. "Experimental Governance: Capacity and Legitimacy in Local Governments." PhD Thesis, KTH Royal Institute of Technology.
- Ettelt, S., N. Mays and P. Allen. 2015a. "Policy Experiments: Investigating Effectiveness or Confirming Direction?" *Evaluation* 21(3), 292-307.
- Ettelt, S., N. Mays and P. Allen. 2015b. "The Multiple Purposes of Policy Piloting and Their Consequences: Three Examples from National Health and Social Care Policy in England." *Journal of Social Policy* 44(2), 319-337.
- Gillespie, S. 2004. "Scaling-up Community-Driven Development: A Synthesis of Experience." FCND Discussion Paper 181.
- Grin, J. 2020. "'Doing' System Innovations from within the Heart of the Regime." *Journal of Environmental Policy & Planning* 22(5), 682-694.
- Haynes, L., B. Goldacre and D. Torgerson. 2012. "Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials." *SSRN Electronic Journal*, ssrn.2131581.
- Head, B. 2010. "Reconsidering Evidence-Based Policy: Key Issues and Challenges." *Policy and Society* 29(2), 77-94.
- Hong, S. and S. Lee. 2018. "Adaptive Governance, Status Quo Bias, and Political Competition: Why the Sharing Economy is Welcome in Some Cities but not in others." *Government Information Quarterly* 35(2), 283-290.
- Howlett, M. and C. Leong. 2022. "The 'Inherent Vices' of Policy Design: Uncertainty, Maliciousness, and Noncompliance." *Risk Analysis* 42, 920-930.
- Hughes, S., S. Yordi and L. Besco. 2020. "The Role of Pilot Projects in Urban Climate Change Policy Innovation." *Policy Studies Journal* 48(2), 271-297.
- Huitema, D. et al. 2018. "Policy Experiments: Core Concepts, Political Dynamics and Impacts." *Policy Sciences* 51(2), 143-159.
- ILO. 2018. "Towards a Stronger Kinofelis: Recommendations for the Strengthening of Future Phases." ILO Report.
- Jacobson, M. and T. Azzam. 2018. "The Effects of Stakeholder Involvement on Perceptions of an Evaluation's Credibility." *Evaluation and Program Planning* 68, 64-73.
- Jessoula, M. and M. Natili. 2020. "Explaining Italian 'Exceptionalism' and its End: Minimum Income from Neglect to Hyperpoliticization." *Social Policy and Administration* 54, 599-613.
- Jowell, R. 2003. "Trying it Out: the Role of 'Pilots' in Policy-Making: Report of a Review of Government Pilots." London, UK: Cabinet Office, Strategy Unit.
- Ko, K. and K. Shin. 2017. "How Asian Countries Understand Policy Experiment as Policy Pilots?" *Asian Journal of Political Science* 25(3), 253-265.
- Koon, A. D. et al. 2020. "A Scoping Review of the Uses and Institutionalisation of Knowledge for Health Policy in Low- and Middle-income Countries." *Health Research Policy and System* 18, 7.
- Lam, D. P. M. et al. 2020. "Scaling the Impact of Sustainability Initiatives: A Typology of Amplification Processes." *Urban Transformations* 2, 3.
- Lehner, L. and M. Kasy. 2024. "Employing the Unemployed of Marienthal: Evaluation of a Guaranteed Job Program," IZA Discussion Paper 16088.
- Leino, H. and M. Åkerman. 2021. "The Politics of Making Finland an Experimenting Nation." *Critical Policy Studies* 16(4), 441-459.
- Leong, C. and M. Howlett. 2022. "Policy Learning, Policy Failure, and the Mitigation of Policy Risks: Re-Thinking the Lessons of Policy Success and Failure." *Administration & Society* 54(7), 1379-1401.
- Lewis, J., M. McGann and E. Blomkamp. 2020. "When Design Meets Power: Design Thinking, Public Sector Innovation and the Politics of Policymaking." *Policy & Politics* 48(1), 111-130.
- Liarde, I., J. I. Criado and L. Alcaide-Muñoz. 2024. "Exploring Barriers to Innovation in Public Administration: An Empirical Study of the Local Layer of Government." *IJPADA* 11(1), 1-23.

- Maddison, S. 2012. "Evidence and Contestation in the Indigenous Policy Domain: Voice, Ideology and Institutional Inequality." *Australian Journal of Public Administration* 71, 269-277.
- Markowitsch J. and Á. Scharle. 2024. "Towards Zero Long-term Unemployment in the EU: Job Guarantees and other Innovative Approaches." Brussels: European Commission. Report.
- Marston, G. and R. Watts. 2003. "Tampering with the Evidence: A Critical Appraisal of Evidence-Based Policymaking." *Australian Review of Public Affairs* 3(3), 143-163.
- Martin, S. and I. Sanderson. 1999. "Evaluating Public Policy Experiments: Measuring Outcomes, Monitoring Processes or Managing Pilots?" *Evaluation* 5(3), 245-258.
- Mattocks, K. 2021. "Policy Experimentation and Policy Learning in Canadian Cultural Policy." *Policy Sciences* 54(4), 891-909.
- Mauw, T., S. Smith and J. Torrens. 2023. "Sustainability Transitions in Los Angeles' Water System: The Ambivalent Role of Incumbents in Urban Experimentation." *Journal of Environmental Policy & Planning* 25(4), 368-385.
- McConnell, A. 2010. "Policy Success, Policy Failure and Grey Areas in-Between." *Journal of Public Policy* 30(3), 345-362.
- McConnell, A., L. Grealy and T. Lea. 2020. "Policy Success for Whom? A Framework for Analysis." *Policy Sciences* 53(4), 589-608.
- McFadgen, B. and D. Huitema. 2017. "Are all Experiments Created Equal? A Framework for Analysis of the Learning Potential of Policy Experiments in Environmental Governance." *Journal of Environmental Planning and Management* 60(10), 1765-1784.
- McFadgen, B. and D. Huitema. 2018. "Experimentation at the Interface of Science and Policy: A Multi-Case Analysis of How Policy Experiments Influence Political Decision-Makers." *Policy Sciences* 51, 161-187.
- Muralidharan, K. and P. Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31(4), 103-124.
- Nair, S. and M. Howlett. 2016. "Meaning and Power in the Design and Development of Policy Experiments." *Futures* 76, 67-74.
- Nesti, G., M. Cittadini and M. Bassoli. 2025. "The Politics of Piloting. The Case of Minimum Income Schemes in European Cities." *Review of Policy Research* 42, 291-309.
- OECD. 2020. "Building Capacity for Evidence-Informed Policymaking: Lessons from Country Experiences." *OECD Public Governance Reviews*. Paris: OECD Publishing.
- Parth, A. M. and J. Nyby. 2022. "Welfare Experiments as Tools for Evidence-Based Policy Making? The Political Debate on Twitter about the Basic Income Trial in Finland." *Policy Studies* 43(4), 772-790.
- Partners for Health Reform Plus. 2004. *The Role of Pilot Programs: Approaches to Health Systems Strengthening*. Bethesda, MD: PHRPlus, Abt Associates.
- Patton, M. 2003. "Utilization-Focused Evaluation." In T. Kellaghan and D. Stufflebeam (eds). *International Handbook of Educational Evaluation*. Dordrecht: Springer, 223-242.
- Pearce, W. and Raman, S. 2014. The New Randomised Controlled Trials Movement in Public Policy: Challenges of Epistemic Governance. *Policy Sciences* 47(4), 387-402.
- Perkiö, J. 2020. "Legitimising a radical policy idea: Framing basic income as a boost to labour market activity". *Policy & Politics*, 48(2), 277-293.
- Raudla, R. et al. 2023. "To Experiment or not to Experiment in Tax Policy?" *Halduskultuur*, 23(1), 27-48.
- Raudla, R. et al. 2024. "It is About Time! Exploring the Clashing Timeframes of Politics and Public Policy Experiments." *Perspectives on Public Management and Governance* 7 (4), 137-148.
- Raudla, R. et al. 2025a. "Budgeting for Innovation: How Funding Strategies Enable or Constrain Public Sector Experimentation." *Public Budgeting & Finance*, 45(4), 28-40.
- Raudla, R. et al. 2025b. "The Politics of Experimental Policymaking: The Influence of Blame Avoidance and Credit Claiming." *Policy Sciences*, 58, 245-266.

- Reymond, Ph. et al. 2018. "Scaling-up Small-Scale Wastewater Treatment Systems in Low- and Middle-Income Countries: An Analysis of Challenges and Ways Forward Through the Case of Egypt." *Utilities Policy* 52, 13–21.
- Reymond, Ph., R. Chandragiri and L. Ulrich. 2020. "Governance Arrangements for the Scaling Up of Small-Scale Wastewater Treatment and Reuse Systems – Lessons from India." *Frontiers in Environmental Science* 8, 72.
- Roseland D, F. Lawrenz and M. Thao. 2015. "The Relationship Between Involvement in and Use of Evaluation in Multi-Site Evaluations," *Evaluation and Program Planning* 48, 75–82.
- Ruggeri, K. 2025. "Assessing Evidence Based on Scale can be a Useful Predictor of Policy Outcomes." *Policy Sciences* 58, 179–188.
- Ruggeri, K. et al. 2020. "Standards for Evidence in Policy Decision-Making." *Nature Research: Social and Behavioural Sciences*, 39905.
- Sanderson, I. 2002. "Evaluation, Policy Learning and Evidence-Based Policymaking." *Public Administration* 80(1), 1–22.
- Simmons, R., P. Fajans and L. Ghiron. [2007]. "Scaling-up Health Service Delivery: From Pilot Innovations to Policies and Programmes." Geneva: WHO.
- Sondal, J., A. Hellquist and B. Balfors. 2024. "From Pilot to Practice: Navigating Pre-Requisites for Up-Scaling Sustainable Urban Solutions." *Urban Transformations* 6, 7.
- Spicer, N. et al. 2014. "'Scaling-up is a Craft not a Science': Catalysing Scale-up of Health Innovations in Ethiopia, India and Nigeria." *Social Science & Medicine* 121, 30–38.
- Stoker, G. 2010. "Exploring the Promise of Experimentation in Political Science: Micro-Foundational Insights and Policy Relevance." *Political Studies* 58(2), 300–319.
- Stoker, G. and M. Evans. 2016. Evidence-Based Policy Making and Social Science. In G. Stoker and M. Evans (eds). *Evidence-Based Policy Making in the Social Sciences*. Bristol: Policy Press, 15–28.
- Strassheim, H. and P. Kettunen. 2014. "When Does Evidence-Based Policy Turn into Policy-Based Evidence? Configurations, Contexts and Mechanisms." *Evidence & Policy* 10(2), 259–277.
- Tcherneva, P. 2019. "The Federal Job Guarantee: Prevention, Not Just a Cure." *Challenge* 62(4), 253–272.
- Tcherneva, P. and A. Lalucq. 2022. "A Job Guarantee for Europe: Economic Policy Innovation for Stability, Security and Sustainability." Brussels: Foundation for European Progressive Studies (FEPS). Policy Brief.
- Torrens, J. and T. von Wirth. 2021. "Experimentation or Projectification of Urban Change? A Critical Appraisal and Three Steps Forward." *Urban Transformations* 3(1), 8.
- Torrens, J., P. Johnstone and J. Schot. 2018. "Unpacking the Formation of Favourable Environments for Urban Experimentation: The Case of the Bristol Energy Scene." *Sustainability* 10(3), 879.
- Turnheim, B. and B. Sovacool. 2020. "Forever Stuck in Old Ways? Pluralising Incumbencies in Sustainability Transitions." *Environmental Innovation and Societal Transitions* 35, 180–184.
- van der Heijden, J. 2014. "Experimentation in Policy Design: Insights from the Building Sector." *Policy Sciences* 47, 249–266.
- Van Parijs, P. and Y. Vanderborght. 2017. "Basic Income. A Radical Proposal for a Free Society and a Sane Economy". Cambridge, MA: Harvard University Press.
- Varazzani, C. et al. 2023. "Seven Routes to Experimentation in Policymaking: A Guide to Applied Behavioural Science Methods", *OECD Working Papers*, 64.
- Verkerk, J. and A. Van Buuren. 2017. "Developing Collaborative Capacity in Pilot Projects Lessons from Three Dutch Flood Risk Management Experiments." *Journal of Cleaner Production*, 169, 225–233.
- Vetulani-Cęgiel, A. and A. K. Cianciara. 2025. "Informational Lobbying in a Polish-German Comparative Perspective: Uncovering Perceptions of Political Relevance of Expert Knowledge and Lobbying Tactics." *Journal of Contemporary European Studies*, 33(4), 1349–1371.

- Vreugdenhil, H. and P. Rault. 2010. "Pilot Projects for Evidence-Based Policy Making: Three Pilot Projects in the Rhine Basin." *German Policy Studies* 6(2), 1-42.
- Vreugdenhil, H. et al. 2009. "The Next Step in Policy Transitions: Diffusion of Pilot Projects." *IRSPM XII Conference Paper*, 1-25.
- Webber, S. and C. Prouse. 2018. "The New Gold Standard: The Rise of Randomized Control Trials and Experimental Development." *Economic Geography* 94(2), 166-187.
- Werner, P. and A. Riedl. 2019. "The Role of Experiments for Policy Design." In A. Schram and A. Ule (eds). *Handbook of Research Methods and Applications in Experimental Economics*. Cheltenham: Edward Elgar, 295-311.
- Wesselink, A., H. Colebatch and W. Pearce. 2014. "Evidence and Policy: Discourses, Meanings and Practices." *Policy Sciences* 47, 339-344.
- Zhu, X. and H. Zhao. 2021. "Experimentalist Governance with Interactive Central-Local Relations: Making New Pension Policies in China." *Policy Studies Journal* 49(1), 13-36.
- Zurbruggen, C. and M. Lago. 2019. "An Experimental Evaluation Tool for the Public Innovation Lab of the Uruguayan Government." *Evidence & Policy* 15(3), 437-451.

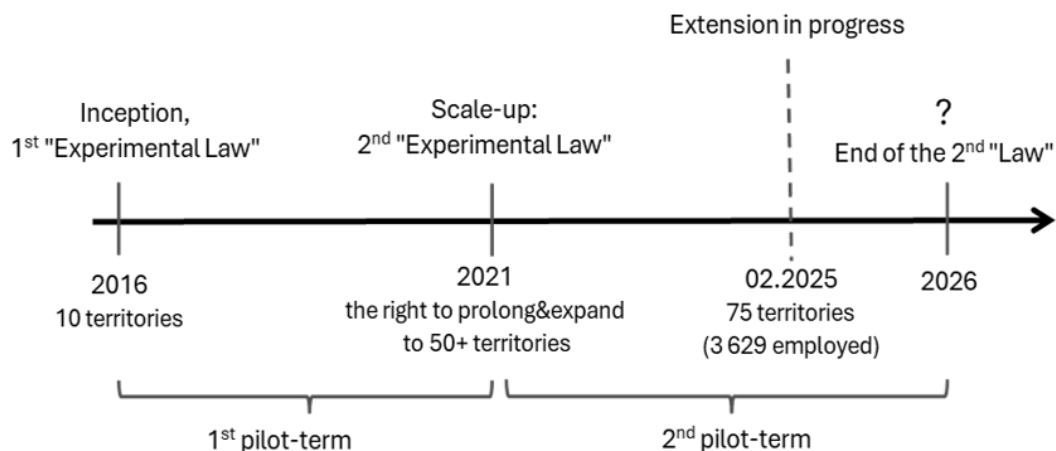
Appendices

Appendix 1. Anonymised list of interviewees

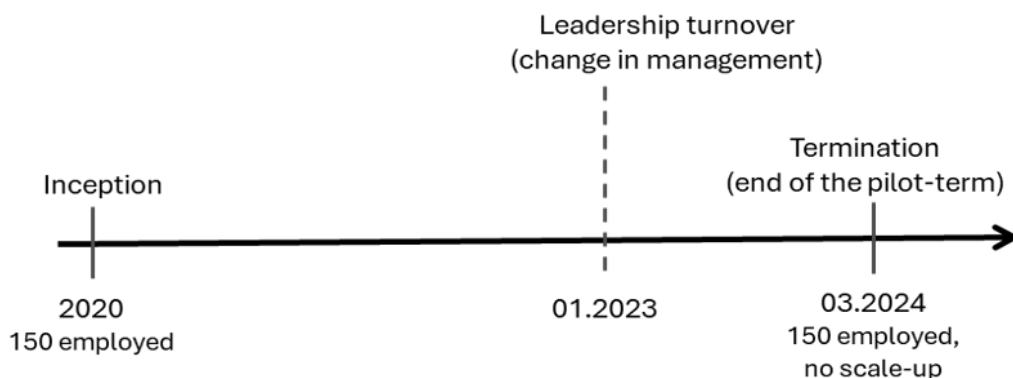
Reference	Role in the program	Interview date
TZCLD, France		
Fr1	National Research & Evaluation Monitoring, Lead	02.06.2023
Fr2	National Management, Vice-President	28.06.2023
Fr3	Funding Management, Board of Directors	01.09.2023
Fr4	National Management & Implementation, Lead	05.10.2023
Fr5	Local Management & Implementation, Lead	19.10.2023
MAGMA, Austria		
Au1	Research/Evaluation	30.05.2023
Au2	Research/Evaluation	07.06. 2023
Au3	Research/Evaluation	28.06.2023
Au4	Implementation, Lead	29.06.2023
Au5	Implementation, Lead	27.07. 2023
Au6	Implementation, Executive	27.07.2023
Au7	Initiator Representative	12.10.2023
Kinofelis, Greece		
Gr1	Initiator	26.10.2023
Gr2	Design/Evaluation, Remote Executive	13.11.2023
Gr3	Research	15.11.2023
Gr4	Design, Implementation, Evaluation: On-site Coordination, Lead	16.11.2023
Gr5	Design, Implementation, Evaluation: Remote Coordination, Lead	08.12.2023
Gr6	Design and Technical Advice	20.01.2024

Appendix 2. Program timelines

TZCLD, France



MAGMA, Austria



Kinofelis, Greece

