Article

Evaluation
1–17
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1356389021997848
journals.sagepub.com/home/evi
SAGE

Tensions of evaluating innovation in a living lab: Moving beyond actionable knowledge production

Rianne Dekker
Karin Geuijen
Utrecht University, The Netherlands

Caroline Oliver
University College London, UK

Abstract

Generative experimentation is increasingly used in public policymaking, especially in response to wicked policy problems. A policy solution is refined within its context and informed by feedback from its users. Studies reporting on these approaches, however, rarely consider the role of evaluation and the nature and goals of knowledge produced. This article addresses evaluation in such contexts. We present a case study of a living lab that combined theory-driven and developmental evaluation, and, responding to contradictory pressures, aimed to generate both actionable and academic knowledge to improve asylum seeker reception. We describe how we addressed these diverging demands and the resulting tensions in a politically charged and substantively insecure policy context. We conclude that evaluation should be an explicit part of the broader design concept, and while generative experimenting can produce actionable learning, evaluation should also aim for academic learning, in a manner that is both democratic and robust.

Keywords

developmental evaluation, generative experimenting, living lab, policy design, theory-driven evaluation

Corresponding author:
Rianne Dekker, School of Governance, Utrecht University, Bijlhouwerstraat 6, 3511 ZC Utrecht, The Netherlands.
Email: r.dekker1@uu.nl

Introduction

For wicked policy problems of 21st century, such as climate change and migration, policy actors find no agreement on the nature of the problem and there are no ‘off the shelf’ solutions ready to be implemented. As a result, public policymaking is moving from traditional ways of planning and implementation towards generative experimenting (Considine, 2012; Kimbell and Bailey, 2017). Here, the goal is to address a problem within its context whereby the policy solution (an idea, innovation, design, policy, programme, etc.) is generated and iteratively refined based on co-creation and continuous feedback from its users (Ansell and Bartenberger, 2016). ‘Living labs’ (Bergvall-Kåreborn and Ståhlbröst, 2009) and ‘design experiments’ (Stoker and John, 2009) are prominent examples that are gaining popularity in public policymaking.

Studies reporting on living labs and design experiments tend to divert more attention to the design process, however, than to the method of evaluation. ‘Labbing’ is sometimes accused of using ‘*quick and dirty methodologies*’ (Tõnurist et al., 2017: 20) to evaluate whether the policy solution works. Similarly, Milley et al. (2018) conclude that few social innovation studies make explicit links to evaluation capacity building. In evaluation studies, by contrast, there is a growing body of knowledge of developmental and theory-driven evaluation research focusing on emerging solutions to messy problems in contexts of complexity. However, these have only scarcely been connected to generative experimentation. Critics argue that that though gaining favour, there is scant evidence regarding the outcomes of co-creation processes and whether these deliver more efficient and effective services, programmes, or policies than less co-creative and user-centred approaches to policy design. A systematic review by Voorberg et al. (2015: 1333) found that most studies of co-creation and co-production identify the contexts and factors which influence the success of the process, but that hardly any attention is paid to the outcomes or generalizability of factors.

Given that there are few studies of evaluating generative policymaking, it remains unclear how evaluation should be embedded in design approaches for public policy and what types of knowledge it produces. Of the repertoire available from evaluation science, evaluation methods differ in the extent to which their aim is to generate lessons for practice, or generalizable evidence-based knowledge on the effectiveness of policy interventions (Leeuw and Donaldson, 2015; Rey et al., 2014). Evaluation has a long tradition of making summative judgements and producing generalizable *academic knowledge*, which may be particularly relevant to the high stakes decision-making of wicked policy problems. Yet equally, evaluation can generate *actionable knowledge* for practice. This is especially the case in developmental evaluation whereby the evaluators support an intervention’s development as it unfolds in real time (Rey et al., 2014). This approach is compatible with initiatives whose focus is on ‘trying things out’ (Patton, 2011: 7), as common in generative experimentation. Theory-driven evaluation is another approach in which evaluators engage with practitioners to construct a theory of change on the effects of a programme in its context (Blamey and Mackenzie, 2007). These methods are best known for producing knowledge for practice. Yet, insight in how these evaluation approaches support different types of learning through generative policymaking in the context of a wicked policy problems is lacking.

This article engages with this debate, by reporting on the experiences of evaluation in a ‘living lab’ in the Dutch city of Utrecht. Here, public and private stakeholders aimed to develop a new concept of asylum seeker reception, addressing head on a ‘wicked policy problem’ which is politically contested and substantively complex to solve. We draw attention to the inherent tensions we experienced as evaluators between on one hand supporting innovation by contributing to actionable knowledge-building, and also harnessing the potentials of producing academic knowledge for questions of merit and scalability demanded of by the funder, broader public and the project management. Within the substantively insecure and politically charged context of a wicked policy problem, there was also a responsibility to produce evidence on whether the initiative had broader merit. This article reflects on this issue with the following research question: *How does developmental and theory-driven evaluation support generative experimentation in a living lab by balancing dual evaluative aims (of producing actionable and academic knowledge) in the context of a wicked problem?*

In the following sections of this article, we first introduce the different types of generative experimentation emerging in public policymaking and the relevance of embedded evaluation models, especially developmental evaluation and theory-driven evaluation. In the next section, we introduce the issue, our case and methodology. This is followed by an account of the evaluation strategy implemented in the living lab and explanation of why it was appropriate to the context of this wicked policy issue. We then present how this evaluation strategy helped yield actionable, but also academic knowledge. Finally, we consider the tensions therein of applying both orientations, but nevertheless conclude that evaluation must contribute to the design by informing not only actionable, but also academic learning. This can be done by simultaneously 'democratizing' theory-driven evaluation and 'evidencing' developmental evaluation.

Generative experimenting in public policymaking and evaluation

Various types of generative experiments are used in public policymaking, including 'living labs', 'design experiments' and 'innovation labs'. They use design methods to analyse problems from different angles and develop, test and improve prototypes to work in a particular context (Stoker and John, 2009; Tõnurist et al., 2017). In living labs, user-needs are central, and collaborative learning is undertaken by users, stakeholders and researchers in a real-life environment (Bergvall-Kåreborn and Ståhlbröst, 2009; Følstad, 2008). Originating from the field of information and communication technology (ICT) innovation, living labs are increasingly used for social and public innovation (Dekker et al., 2019). Living labs use a process of generating and iteratively refining a policy solution in a real-world setting. They follow an abductive logic of coming up with provisional hypotheses as a basis for developing prototypes, which *may* prove a sufficient solution to the problem (Kolko, 2010). This includes bringing forward solutions that were not envisioned at the start of the project (Gascó, 2017). Users co-design the concept, while after each design cycle, the solution is evaluated based on their experiences. This is referred to as 'user-driven innovation' (De Moor et al., 2010).

Similar to design experiments and innovation labs, living labs require evaluators to immerse themselves in the real-world policy context in which the design takes place (McGann et al., 2018). Evaluators work closely with the participants and partners in the face of unexpected turns and unplanned outcomes. This requirement resonates with recent evaluation approaches of policy responses to challenging problems, characterized by innovation, contextual influences and complexity. Developmental evaluation is one of these approaches, which is increasingly applied to complex innovation environments where the goals and paths towards these goals are evolving (Lawrence et al., 2018; Patton, 2011). The primary focus of developmental evaluation is on adaptive learning rather than accountability, as evaluators recognize that the very notion of ‘what works’ itself is subject to change under conditions of complexity (Patton, 2011: 188). The approach is flexible, with new measures and monitoring mechanisms evolving as understanding of the situation deepens and the project’s goals emerge, with the evaluator intervening in the innovation process by providing intermediate and real-time feedback to inform further development. Developmental evaluation is primarily ‘utilization-focused’: concerned with generating actionable knowledge and fostering use of evaluation results, rather than providing summative judgement, assessing causality and producing accountability reports (Contandriopoulos and Brousselle, 2012; Rey et al., 2014).

Theory-driven evaluation is a more evidence-based approach prioritizing academic research to demonstrate outcomes and reasons why they were or were not achieved. Theory of change has particular traits that work well in conditions of complexity. In this approach, the evaluator reconstructs the assumptions of practitioners in a project on how and why a solution will work (Weiss, 1997). Assumptions are tested by studying the links between activities, outcomes, and contexts of the solution (Connell and Kubisch, 1998; MacKenzie and Blamey, 2005). Rather than seeing programmes as unified entities through which recipients are processed in a linear fashion, theory-driven approaches recognize that people are subjected to policy programmes in different ways under different circumstances (Blamey and Mackenzie, 2007). Ultimately, theory of change evaluation aims to produce and integrate both actionable and academic knowledge in producing an understanding of whether and why outcomes were met (Weiss, 1997).

These two models provide useful starting points for considering further the role and value of evaluation of generative experimenting in public policymaking, where programmes and systems in which they operate are complex and do not yet include the ‘finished product’. Especially in developmental evaluation, the utilization focus, primacy of the user in design and evaluation and reflexivity demanded of the evaluator fit the objectives and values of co-creation and innovation taking place in living labs. The theory of change approach promotes attention for diverging ideas on workable solutions, but still nevertheless builds some pathway towards anticipated, pre-defined goals. Theory of change also seeks understanding of the context in which the policy is implemented to explain differences in effectiveness and user-experience.

The evaluation approach we took combined developmental evaluation with theory of change evaluation. The aim of evaluation was to support innovation, but a completely ‘outcome free context’ was not possible, nor ultimately desirable if genuine lessons were to be learned. We consider to what extent the claims for these approaches in producing knowledge for practice and generalizable knowledge might be more ideal than real, especially when undertaken in substantively insecure and politically charged real-world context of wicked policy problems. Earlier research by Rey et al. (2014) into developmental evaluation distinguished three tensions in producing both actionable and academic knowledge: (1) linking research and evaluative objectives, (2) the dual role of researcher and consultant and (3) the temporality of the process. We add to this literature by considering the tensions in the context of a living lab addressing a wicked policy problem and reflecting on how we dealt with these.

The case: A living lab to innovate a response to the wicked policy issue of asylum seeker reception

The Utrecht Refugee Launchpad (U-RLP) was a living lab situated in the city of Utrecht, the Netherlands. In aiming to innovate asylum seeker reception, it engaged with a wicked policy problem: an issue that does not lend itself to traditional, mainstream public problem-solving, and where there is much uncertainty on the best solution due to incomplete and contradictory knowledge (Alford and Head, 2017). Wicked problems are also characterized by political and institutional difficulties, as they invoke potentially conflicting values, and strong, opposing political ideas among multiple parties. This is certainly the case in the policy area of reception of asylum seekers. Disputes abound among stakeholders at the local, national, as well as on the European level about the purpose of reception: to protect asylum seekers' human rights or to protect national (economic and cultural) interests (Geuijen et al., 2017; Noordegraaf et al., 2019). Solving the problem is difficult when there is large-scale disagreement about whether, and how far policies should help asylum seekers to integrate, a complicated issue as while some are allowed to stay in the country, others will have their application rejected and will either have to leave the country voluntarily or be deported.

In the Netherlands, asylum seekers spend their first months, and even up to some years in asylum seeker centres (ASCs). These centres provide housing during the period when their asylum request is being assessed and before those who have a residence permit can move into regular housing. Similar to other European countries, reception in the Netherlands is designed to be 'basic but humane' with little possibilities for asylum seekers to integrate into Dutch society (Advisory Committee on Migration Affairs (ACVZ), 2013). However, studies show that time in an ASC can lead to worsened well-being of asylum seekers and prolonged unemployment after departing the ASC (Engbersen et al., 2015).

Developing alternative forms of asylum seeker reception to solve these issues has proven substantively difficult and politically contested, with some arguing for better facilities and services to lead to early integration, and others arguing for more basic provisions to discourage arrival and allow for easier deportation. It is also challenging to prove the worth of alternatives, due to difficulties in providing 'hard' results of interventions to improve outcomes. For various reasons, including the effects of trauma and disrupted educational trajectories, refugees' social and economic integration is protracted and research across Europe shows that it can take 10–15 years for a refugee to reach parity in employment with other migrants arriving at the same time (Connor, 2010).

As is common to living labs, U-RLP was initiated and led by a consortium of public and private partners. The consortium gained European funding through the Urban Innovative Actions (UIA) programme, a funding scheme designed to provide urban areas throughout the European Union (EU) with resources to experiment and test new and unproven solutions to solve urban challenges. A partnership consisting of the city of Utrecht, various NGOs and knowledge institutes set out to advance a new concept in asylum seeker reception with its intended beneficiaries (asylum seekers and locals) within the setting of an ASC. The concept was framed around two ambitious pillars of 'co-housing' and 'co-learning' of asylum seekers with local residents.

While the traditional Dutch model of reception of asylum seekers consists of large-scale facilities housing several hundreds of asylum seekers, isolated from Dutch communities, this specific ASC housed a lower number of up to 400 asylum seekers together with 38 local youngsters, who lived in a building adjacent to the ASC. The combination was chosen based on programme managers' assumption that people who are 18–30 years old are in a more flexible and open-minded phase of their lives, and open to establishing contacts with strangers as well as initiating activities. Moreover, because of the severe housing shortage in Dutch cities, it was assumed that students and young people starting in employment would be glad to find housing, especially with a reduced rent offered as an additional incentive to participate.

In the complex, the asylum seekers and young people shared facilities including a common living room and classrooms (referred to as the ‘incubator space’), a kitchen and an outside terrace. In traditional ASCs, asylum seekers are able to do only very limited activities like sports and voluntary work such as cleaning the facilities, which can provoke feelings of uselessness and boredom, and ill-prepare them for the labour market (ACVZ, 2013; Engbersen et al., 2015). Project partners in the living lab instead delivered an educational programme in Business English and Entrepreneurship for asylum seekers and local people from the neighbourhood. As the intervention unfolded, many additional activities were added ranging from visits to museums, collaborative cooking and social events, to basic computer courses, ‘language cafes’ providing informal settings to practice the Dutch language with students and other volunteers, and professional networking events.

Beyond some basic characteristics that were set out in the project plan, U-RLP was expected to unfold and take shape in its own way. There was budget and opportunity for the asylum seekers and locals to develop activities that would support integration and well-being. The ASC complex and the surrounding neighbourhood of Overvecht provided the ‘testbed’ for designing and testing of these newly co-created interventions. Asylum seekers and locals were seen as beneficiaries of the project: the inclusive approach was expected to benefit the asylum seekers starting a new future as well as the local residents of the vulnerable neighbourhood of Overvecht where the ASC was situated. In the next section, we explain how the evaluation responded.

The method: Combining developmental and theory of change evaluation in a living lab

The funding for the project stipulated an evaluation of the U-RLP living lab, and the academic research team was invited into the partnership prior to the project’s inception. As such, we had exclusive access to the research site from the outset to collect data on the ongoing process of innovation and its effects on different groups involved in the project. We held recorded interviews and meetings with representatives of the project steering group, including project leaders from the municipal council, a housing corporation, a NGO working with refugees, a social enterprise and higher education institutions. As shown in Table 1, we used multiple research methods to learn about participant experiences, including quantitative methods of monitoring their participation, as well as surveys and qualitative methods of interviewing and ethnographic observation to understand people’s experiences of the project (Table 1). This ‘bricolage’ of methods is common in developmental evaluation (Patton, 2011: 264). Findings are based on our research throughout the full timeframe of the project (February 2017–October 2018) and the year afterwards when the concept was transferred to another ASC in Utrecht. Our evaluation strategy and results were documented in notes and memo’s, minutes of team meetings and products of the evaluation research including an interim and final project report.

The perspective of the ‘user’ was central to our evaluation, which included the perspectives of the asylum seekers and local youngsters living at the complex and locals from the neighbourhood as well as multiple project partners. The first step for the evaluation team was to help articulate how ‘co-housing’ and ‘co-learning’ would lead to change. We interviewed the project partners about their role in the project and held a workshop to elicit a theory of change

Table 1. Overview of data collection.

Type	Sample	Data collection
Quantitative data on courses and activities	Project partners shared data on 21 process indicators. These include participation in courses and activities, group composition, completion rates, etc.	Registration by project partners throughout the project
Face-to-face surveys in the neighbourhood	Simple random sample of all 6552 addresses in the area directly around the ASC	Cross-sectionally in two waves: autumn 2017, *N* = 304 and autumn 2018, *N* = 277
Online surveys among the youth tenants	Census sample of population 38 youngsters	Cross-sectionally in two waves: winter 2017, *N* = 23 and winter 2018, *N* = 19
Online intake assessments by asylum seekers who started participation in the courses	Self-selection among 558 adult asylum seekers who chose to actively participate in the programme	*N* = 150 asylum seekers who signed consent to share their results
Course evaluation surveys at the end of each 8-week course	Census sample among participants in English classes and Business incubation programme	*N* = 206 responses collected at the end of each course
Semi-structured face-to-face interviews with asylum seekers	Convenience sampling of asylum seekers participating in the programme and some who did not. We strived for repeated interviews, but this was not possible for all	*N* = 83 interviews throughout the project, of which *N* = 21 repeated
Semi-structured, face-to-face interviews with young tenants	Convenience sampling among 53 youngsters who lived in the building throughout the project	*N* = 19 interviews throughout the project of which *N* = 5 repeated interviews
Semi-structured, face-to-face interviews with locals from the neighbourhood	Recruitment from the general neighbourhood population and those who participated in courses and activities	*N* = 38 interviews throughout the project, and after the two survey waves
Semi-structured, face-to-face interviews with project partners	Interviews with at least one representative from all partner institutes. Recorded interviews were supplemented with records of multiple informal conversations	*N* = 20 formal interviews including four repeat respondents; informally, conversations maintained throughout the project
Ethnographic research	Participant observation at a variety of occasions and during steering group meetings. Field notes were taken. Project documents including the project plan, agendas and minutes of meetings and project communications were also analysed	On a variety of occasions during the project and after the closure of the ASC
Media reports on U-RLP	Dutch and English newspaper sources between beginning of January 2016 and March 2019 retrieved through Nexis Uni and TV items collected through the database of the Netherlands Institute for sound and vision	Full sample of N = 307 newspaper articles and TV items

ASC: asylum seeker centre; U-RLP: Utrecht Refugee Launchpad.

that they collectively held about how the project would work. Three goals were identified: first better well-being for asylum seekers during their stay in the ASC; second, better preparation for labour market integration after receiving a residence permit, and third, better relations with the surrounding neighbourhood (as locals in the district of Overvecht had reacted negatively on the local authority's plans to open the ASC). Translating the rather 'aspirational language' of the project application into measurable concepts was challenging (cf. Lawrence et al., 2018). Benefits were expected in three areas: (1) relations of asylum seekers with the neighbourhood, (2) skills and labour market prospects of the asylum seekers and (3) well-being and self-efficacy of the asylum seekers. There were noticeable differences in partners' views on the project and proposed project outcomes, revealing again the complexity of the problem and the conflicting values held about the topic.

The ASC was home to asylum seekers and local youngsters for several months, up to a year and a half. Through the multiple methods, we tracked their lived experiences, adopting a longitudinal approach while monitoring changes in the concept. Recognizing that this was a fluid and unfolding initiative, we monitored what activities took place in the project as part of, and also beyond this initial concept collecting data continuously and cross-sectionally (Table 1).

As is consistent with developmental evaluation, we also aimed to provide timely, utilization-focused feedback to the project partners, throughout the project – primarily at the bimonthly steering group meetings of the project partnership. The evaluation was a recurring agenda point, where we would report on our activities and early findings. We would contribute to discussions, providing insights from the data collected to inform decision-making. Eighteen months into the project, we presented findings and recommendations on the full concept in an interim report and policy brief. This fed into the second phase of the project and transfer of the concept to another ASC in the city, where many of the lessons learned informed decision-making. Based on this feedback, and also based on the partners' and participants' own experiences, aspects of the concept were adapted along the way. This enabled a learning effect to take place throughout the project, to tighten and improve the developing concept. Using developmental evaluation principles, we were reflexive in our approach and explicitly discussed our own impact on the experiment in the evaluation reports.

Knowledge production and its consequences: Actionable and academic knowledge

In the following sections, we present an account of the two different types of knowledge produced through the evaluation and consider how each made a difference for the programme, particularly in the context of addressing the wicked issue of asylum seeker reception.

Actionable knowledge: Supporting innovation and legitimate impact

The evaluation provided actionable knowledge, or utilization-focused knowledge in three main ways. First, our approach supported the focus on collaborative experimenting and learning rather than on accountability, and encouraged inventiveness among the participants and project management. While there were some pre-established targets of what the project would deliver relating to the course programme (8-week courses in Entrepreneurship and English) monitoring showed that some of these targets were met (e.g. the percentage of locals that enrolled), but others (such as the numbers of asylum seekers enrolled) were not. Tracking these numbers was not meant to pass a verdict on whether the project succeeded. Instead, it enabled reflection on why or why not targets were met and helped partners re-align these to the project's aims within its changing context. These discrepancies could be explained as the project ran for a shorter period in practice than envisaged, due to delays beyond the initiative's control in placing asylum seekers in the initiative, and we also fed back that courses did not speak to the needs of all asylum seekers. Some targets lost their urgency or proved counter-effective or overly ambitious over the course of the project. In these cases, intermediate feedback from our evaluation helped inform practice to diverge or retain them. Though initially engaging with this monitoring was seen as onerous, many partners fed back that they valued the exercise, as it offered helpful insights into where their initial plans had changed.

Our approach to evaluation also took unplanned results and newly conceived plans into account. For example, as the project developed, the partnership sought collaboration with multiple other projects and organizations working with asylum seekers in the city of Utrecht. This included 'Welkom in Utrecht', an NGO organizing various types of social activities as well as connecting people to volunteer work, and InclUUsion, a project offering participation in Utrecht University courses. New ideas emerged, such as a redesign of the central meeting space of the initiative by the youngsters and asylum seekers into a 'living room'. This was done to give it into more of a relaxed atmosphere instead of the trendy 'office' atmosphere it had initially, and also to attract more locals from the neighbourhood – stimulating mutual contact. These examples indicate how both the project partners and participants felt encouraged to develop the concept beyond the initial idea. Our evaluation recognized changes to the initial plan and measured their effects, still broadly within the outcomes of enhanced well-being and connection, but with clearer understanding of the steps that led to change. As a result, the evaluation was supportive of change and inventiveness by the participants along with a changing context and demands.

Second, benefits for learning for practice resulted from our position as evaluators: we were part of the steering group and we shared their values and ambition to develop a new concept of asylum seeker reception. We positioned ourselves as critical partners or consultants (cf. Patton, 2011; Rey et al., 2014), but yet who were equally committed to the process of innovation. As a result of this embedded type of evaluation, we noticed that our feedback was taken seriously and there was willingness to adapt the plan accordingly. For example, at the start of the project, the partnership chose to use a 'word of mouth' communication strategy towards the neighbourhood, not to risk stirring public unrest or even hostility to the project. In surveys and interviews among locals, we found that this low-key communication strategy resulted in the fact that large groups of neighbourhood residents in the first year did not yet know about their opportunities to participate. After communicating this in a steering group meeting, the project team developed a more open communication strategy towards the neighbourhood, for example by visibly drawing attention to the centre through banners outside.

As a result, from the early stages of the project, evaluation had real impact on the concept, and therefore, the lives of asylum seekers and locals. By participating in the courses and activities programme, all asylum seeker participants felt that they were using their time more productively in comparison with their stay in other ASCs. There, they experienced more feelings of boredom and depression. They also valued the skills they had learned and gained a new focus on a future in the Netherlands. Locals from the neighbourhood and youngsters living in the project felt that they could make a difference for refugees by participating in the project. They gained themselves too from participating, by learning skills and incubating new business ideas.

Table 2. Overview of evaluation's contributions to actionable and academic knowledge.

Actionable knowledge	Academic knowledge
Intermediate feedback on planned and unplanned aspects of the concept leading to its improvement	Longitudinal measurement of effects of changes in the concept
Embedded evaluation enhancing reflexivity and learning by partners and participants	Real-world setting in which the concept is developed and tested improving ecological validity
User perspective contributing to legitimacy of the concept	Triangulation of methods helps gaining rich insight and internal validity

Third, benefits for practice resulted from the evaluation being rooted in user-experience. Initial plans for an ASC in this neighbourhood were met with public protest. By focusing on both asylum seekers and locals' perspectives, the evaluation generated legitimacy of the concept (cf. McGann et al., 2018). We were asked to present evaluation findings to local and national policymakers and to policymakers from other cities in Europe. Eventually, evidence on the acceptance of the ASC by the neighbourhood and benefits experienced by asylum seekers and locals led to transfer of the concept to another ASC across the city, with evaluation evidence drawn on to adapt the concept to the conditions there, where there was a more accepting local neighbourhood. Yet, at some moments evidence from the evaluation was ignored and political arguments prevailed in decision-making about the project. In particular some of the partners feared that the concept would lose its trademark and thereby political and public support. This mirrors the classical balancing act between maintaining as much of an innovation's evidence base on one hand to be able to prove the innovation's robustness in different contexts, while on the other hand adapting an innovation as much as possible to a new local context to make it fit the new specific local needs (Williams, 2020).

Summarizing, the evaluation approach we implemented in the living lab, supported the production of knowledge for practice in three ways (Table 2): First, the focus of evaluation was on both planned and unplanned aspects of the concept as well as supporting ongoing development through delivering intermediate feedback. Second, as evaluators being part of the project partnership, we developed a trusting relationship with partners and participants which encouraged reflexivity and learning and led to real-time impact of evaluation results. Finally, by taking user-experiences as the basis for evaluation, the concept gained legitimacy even in a politically contested context and was transferred to another ASC.

Academic knowledge: A pragmatist knowledge base for policymaking

Providing actionable knowledge certainly had benefits, but the contextual reality of innovation programming meant that questions of merit and generalizable knowledge did not disappear entirely from the frame. The funders had expected evaluation to deliver evidence on measurable results, while generating understanding of 'what worked and what did not, and why so, what should be done differently' (European Commission, 2016). This was particularly to inform transferability of the practice elsewhere, with many European municipalities keen to learn from the innovation. Likewise, there was much public and political interest within the city government as to what the project could deliver. As such, engaging with questions of effectiveness and merit, even in an early, and less developed stage of a concept, we argue, offers important opportunities to advance academic knowledge and inform practice in ways that are valuable beyond the life of a project. We summarize how we did so below.

First, engaging in longitudinal measurements enabled us to capture changing effects of the concept in comparison with its earlier designs. For example, after changes were made towards a more open communication strategy to the neighbourhood, we observed that the numbers of visitors and participants from the neighbourhood went up. While causality could not be established here (the centre had been present for longer and the offer of courses and activities had changed as well), we were able to produce an account of this change in the concept towards higher participation while considering other plausible explanations for greater engagement. Through a door to door survey and interviews we also learned that the neighbourhood became more knowledgeable of the ASC and that hostility towards the centre was initially lower than commonly understood. This understanding, combined with a deeper integration of partners' theories with academic research on public reception to asylum seeker reception led to an important recommendation that while communication strategies should be respectful of hostility as a public response, they should not be led by the assumption that this was the dominant response.

Second, evaluation in the living lab contributed to academic knowledge rather than only contextual knowledge because of its ecological validity (cf. Shadish et al., 2002). The new concept of asylum seeker reception was developed and trialled in a real-world setting where we could measure the lived experiences of participants with the new concept of asylum seeker reception. For example, we could monitor participation of asylum seekers in courses during a difficult time of legal insecurity rather than in a hypothetical situation and where the broader geopolitical context could be taken into account in explaining some of the outcomes of the project. For example, the EU–Turkey refugee deal in March 2016 and the closing of European borders caused a reduction in the numbers of asylum applications all over Europe. This led to significant delays in arrival of asylum seekers to the ASC. This had effects on momentum of the project experienced by the local youngsters who had already been living at the centre for months, as well as on the people in the neighbourhood. Such effects could not have been demonstrated in an artificial research setting. The living lab approach enabled us to demonstrate the effects of such volatility in real time, and build these explanations into understandings of the 'success' or otherwise of the project.

Third, triangulation of data contributed to internal validity of evaluation results. The combination of methods helped to gain rich insights in the experiences of asylum seekers and locals, but through integrating them we could confront discrepancies between data sources. An example of this was online assessments on well-being and labour market aspirations that asylum seekers completed before entering the course programme. These were offered by an external provider and used by the project partners to advise asylum seekers on what courses to take. They also offered relevant data for evaluation research. However, qualitative interviews among asylum seekers and some partners working with asylum seekers revealed that asylum seekers experienced this assessment as a 'test' on which they had to perform well. They believed there were potential incentives associated with the assessment results (i.e. opportunities in work and education). We, therefore, concluded that this might have led to socially desirable answers and took caution in interpreting the assessment results.

Summarizing, our evaluation approach in the living lab offered three specific advantages for gaining academic knowledge (Table 2). These are gaining evidence on the plausible contribution of the concept on the basis of measured changes over time, building in the real-world setting of in which interventions are tested and ecological validity, and triangulation of research methods and internal validity. In these ways, evaluation contributed to eliciting the merit of parts of the concept and their scalability.

Tensions between academic and actionable learning

Our account thus far provides a useful case study of how evaluating in a living lab produced actionable and academic learning through generative policymaking. However, similar to Rey et al.'s (2014) applications of developmental evaluation, we experienced pressures of being subject to conflicting demands between dual aims of evaluation and research, especially in the wicked policy context of asylum seeker reception. These are summarized in Table 3. In the following, we discuss the limits of actionable knowledge and show some of the tensions that resulted when aiming to produce academic knowledge too.

First, while using multiple moments and types of data collection in the living lab led to more robust research findings that would stand up to external scrutiny, we found that this could overburden participants and project partners. For example, we asked project partners to systematize their initially rather low-key approach to registering participation, requiring more investment of time and energy. Registering characteristics of participants such as gender, age, and country of origin required categorization also created some consternation as partners felt this went against the inclusive spirit of the project. In response to this issue, we made registration as easy as possible and put effort in explaining why some measures were important for learning and reporting.

In deciding what concessions to data collection we could make, we took into account the different bases on which partners and participant groups participated in the living lab: we could ask more of the partners and youngsters who chose to become part of the project. The asylum seekers however were not able to make any choices on where to live and for how long. They were assigned to this ASC by the government agency managing asylum reception. Therefore, we decided it would not be ethical to engage them with an additional questionnaire on top of the lengthy assessment they were already taking upon entering the project. Here, we opted to use the assessment results as a suboptimal, but more ethical compromise.

Second, we experienced that commitment to the aims of the living lab incited pro-innovation bias among the project partners and us as evaluators. This entails a bias towards innovation over the existing situation (cf. Karch et al., 2016). Partners and participants felt that they were part of creating something unique and there was a risk of being sometimes overly content with what it achieved. There were risks that working closely within the team might compromise academic research integrity, especially when evaluators lose the 'speaking truth to power' element of their role (Gamble, 2008 in Rey et al., 2014). Becoming aware of this, we put effort in recruiting less involved participants to incorporate a greater diversity of experiences with the

Table 3. Tensions between gaining academic and actionable knowledge in a living lab.

Tension	Response chosen
Data collection as a burden on participants and partners	Limiting and simplifying methods of data collection by considering the commitment of different participant groups
Pro-innovation bias and research integrity	Recruiting less involved participants, consulting an academic advisory board and gaining a comparative perspective from similar experiments
Timing of evaluation results	Quicker turnaround of reporting for actionable learning and political decision-making than for academic learning
Type of 'proof of concept' evaluation would deliver	Making agreements within the partnership about internal and external reporting

project, instead of getting positive feedback only. Naturally, we ensured anonymity and confidentiality of participants' individual accounts of the project so that they could be critical of the design. Yet this more critical stance affected our positioning within the project team and at times risked us being viewed as 'against' rather than 'with' them.

Maintaining distance from the project was also achieved by subjecting our evaluation design and ongoing work to scrutiny by an academic advisory board, which met annually. This comprised experts in the fields of refugee and migration studies, as well as evaluation. This kept any pro-innovation bias in check, by empowering us to speak back to the project managers on issues where evidence suggested less positive gains, and encouraging us to be more critical in our final report. We also sought exchanges with other projects in the Netherlands and Europe where new forms of asylum seeker reception were being developed and tested. This enabled us to gain a comparative perspective and exposed our own findings to external, critical scrutiny.

Third, there were conflicting demands put on timing of evaluation results being published. For evaluation to have impact on policymaking, timing of results being fed back proved crucial. The practitioners in the project needed timely information to develop solutions for daily problems in the ASC as well as for advising on policy decisions to be taken. In some cases, deadlines were too tight for putting to work rigorous academic knowledge. For example in September 2018, an initial plan to close the ASC by November 2018 was reconsidered by the Alderman. Key in this decision was evidence from the evaluation on the neighbourhood's response to the centre. At that moment, the second round of the neighbourhood survey had only just been rounded up and we could not give any representative data on the neighbourhood's attitude at that point. Instead, we chose to refer to evidence from the first survey that was communicated in the interim report and conveyed the message that we had no reason to expect that this would have changed, on the base of our quick analysis of only some parts of the second survey. For yielding academic knowledge for the final report and publication, we insisted on taking time to gather and to analyse data to be able to provide rigorous conclusions and recommendations. Rey et al. (2014) already concluded that the reflection time required for theoretical analysis is quite incompatible with the rapidity of factual feedback expected in social innovation. Here, we add that this is even more present with the turnaround of political decision-making and unforeseen developments in the wicked problem context.

Finally, we experienced conflict over the type of 'proof of concept' that the evaluation would deliver. The 'refugee crisis' combined with the EU's urban agenda, and the city authority's positive backing, had given the partnership a unique opportunity to start this project in face of public and political opposition. The partnership, therefore, needed to showcase the project and its evaluation. This posed a challenge to our evaluation approach in which improving the concept by learning from error was key. To meet this challenge, we made publication agreements about what feedback would be provided internally and publicly and at what times. Here we ensured that reporting would not harm the project's impact internally or externally. There also was discussion about the level of analysis on which evaluation could draw conclusions. The partnership was keen for us to provide evaluative claims on the concept's overall success. Media reports also made claims that 'the project' worked based on anecdotic evidence form a few key participants. The evaluation, however, led to conclusions about the contribution of *specific aspects* of the concept such as the co-housing or the co-learning parts. This sometimes led to disappointment among partners and outsiders with an interest in the project. The time frame of the project was too short to find long term effects on, for example, well-being and labour market participation. Here we kept our focus on the more granular level of analysis and we would leave broader evaluative claims to others.

Conclusion

'Messy problems require messy solutions' argue Verweij and Thompson (2006) in their case for 'clumsiness' in developing policy solutions. Indeed, in cases of wicked policy problems, policymakers increasingly turn to generative experimenting through living labs and other design approaches as a method to develop and test policy solutions in an abductive fashion. Policies are not planned and implemented top-down but co-created with intended beneficiaries in an open-ended process of innovation. This is an appropriate approach to wicked problems which cannot be solved in the traditional sense of the concept, because of their substantive complexity and political controversy. The goal becomes about making progress rather than finding a solution. However, this can lead to problems when co-creation and co-production is seen as a 'virtue in itself' (Voorberg et al., 2015: 17) and no longer has to justify itself through providing evidence on outcomes or meeting external objectives. As we have argued, there are risks of not embedding and explicating a method of evaluation within generative policymaking. Demonstrating to what extent and how practical alternatives work adds evidence-based knowledge to the substantive complexity which characterizes wicked problems. Researching, explicating, including and evaluating stakeholders' different opinions, assumptions and values helps to reframe the issue and contributes to opening up the political deadlock which characterizes wicked problems.

Based on a case study of a living lab innovating asylum seeker reception, this article explains how evaluation can be applied within generative experimentation. On one hand, it is well placed for producing actionable knowledge through developmental evaluation that informs what needs to change by '*crafting new solutions with people, not just for them'* (Carstensen and Bason, 2012: 6). However, we also show how evaluation which delivers a form of judgement is still called for, even in developing concepts, especially when high stakes questions of wicked policy issues are involved. Here, scrutiny means that an 'outcome free context' of innovation is not realistic, and accountability cannot be completely deferred to a later period beyond an intervention's creative period of development (Patton, 2011). Therefore, over the course of some months, we worked to make explicit the largely implicit theories of change of a multiplicity of partners, making clearer partners' assumptions about how the planned interventions would lead to outcomes on the selected goals. This process of co-creation of the theory of change led to a more robustly substantiated selection of interventions over time. Informed by rigorous research, inferences on effects of the intervention were made, enabling us to assess the contribution that changes to the intervention had made in comparison with earlier versions of the concept, as well as relate the claims to academic work on similar interventions elsewhere (see Befani and Mayne, 2014). The real-world setting in which the concept was tested, improved ecological validity of the findings and triangulation of methods helps gaining rich insight and improve internal validity.

Of course, combining such approaches is not an easy fit. The two approaches differ in their views on the position of evaluation and especially on the need for summative judgement of the policy solution. Developmental evaluation is a bottom-up, participatory approach, empowering partners and users and prioritizing their experiential knowledge. Theory-driven evaluation is a more evidence-based approach prioritizing academic research to demonstrate outcomes and how and why they were or were not achieved. Actively co-creating a robust theory of change and periodically feeding our scientific analysis from following the project development back into this process allowed for decisions to be taken on next steps. Working in this way enabled simultaneously 'democratizing' the theory-driven evaluation approach, and 'evidencing' the developmental evaluation approach. It shows the real possibilities of reconciling generative experimenting and collaborative processes of design thinking in policy development and evaluation on one hand, with evidence-based policymaking and evaluation on the other hand.

A key contribution is making explicit the conflicts that arise from aiming to meet diverging demands of academic and actionable knowledge production. Extending Rey et al.'s (2014) findings, we draw attention to the important tradeoffs that needed to be made in the context of a wicked policy problem. We had to adopt concessions on what we could ask, to avoid overburdening partners and participants in the living lab. We used an advisory board and exploited opportunities for wider comparison with similar project to limit pro-innovation bias. We were constantly under pressures around the timing and type of reporting our evaluation results would deliver. This only confirms as Lewis et al. (2020) conclude, that design processes, when coming into contact with power and politics, face significant challenges. In a volatile political context of wicked policy problems, policy-relevant outcomes need to be presented quickly and judgements focus on whether 'it' worked. For academic knowledge, more time for reflection is required, with evaluative claims focused on parts of the concept, rather than the whole.

We conclude that explicit attention for evaluation, and the merits of different approaches is needed to support knowledge production through living labs and other types of generative experimenting. By implementing proven evaluation strategies, living labs can provide a basis for not only instrumental and democratic actionable knowledge, but also academic learning on the effects of newly developed policy interventions that can lead to better understanding of complex problems and how we can address them.

Acknowledgements

The authors would like to thank the partners, stakeholders and users of the Utrecht Refugee Launch Pad for their openness during the evaluation. They thank their colleague Yousra Rahmouni Elidrissi for helping out with the French translation of the abstract. They also thank the anonymous reviewers of *Evaluation*. Their feedback has made all the difference.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was supported by the European Regional Development Fund through the Urban Innovative Actions Initiative (grant number UIA-01-131-URLP).

ORCID iD

Rianne Dekker https://orcid.org/0000-0001-6460-4223

References

Advisory Committee on Migration Affairs (ACVZ) (2013) *Verloren tijd: Advies over dagbesteding in de opvang voor vreemdelingen*. The Hague: ACVZ.

Alford J and Head B (2017) Wicked and less wicked problems: A typology and a contingency framework. *Policy and Society* 36(3): 397–413.

Ansell CK and Bartenberger M (2016) Varieties of experimentalism. *Ecological Economics* 130: 64–73.

Befani B and Mayne J (2014) Process tracing and contribution analysis: A combined approach to generative causal inference for impact evaluation. *IDS Bulletin* 45(6): 17–36.

Bergvall-Kåreborn B and Ståhlbröst A (2009) Living Lab: An open and citizen-centric approach for innovation. *International Journal of Innovation and Regional Development* 1(4): 356–70.

Blamey A and Mackenzie M (2007) Theories of change and realistic evaluation: Peas in a pod or apples and oranges? *Evaluation* 13(4): 439–55.

Carstensen HV and Bason C (2012) Powering collaborative policy innovation: Can innovation labs help. *The Innovation Journal: The Public Sector Innovation Journal* 17(1): 1–26.

Connell J and Kubisch AC (1998) Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects and problems. In: Fulbright-Anderson K, Kubisch A and Connell J (eds) *New Approaches to Evaluating Community Initiatives, Vol. 2, Theory, Measurement, and Analysis*. Washington, DC: Aspen Institute, 15–44.

Connor P (2010) Explaining the refugee gap: Economic outcomes of refugees versus other immigrants. *Journal of Refugee Studies* 23(3): 377–97.

Considine M (2012) Thinking outside the box? Applying design theory to public policy. *Politics & Policy* 40(4): 704–24.

Contandriopoulos D and Brousselle A (2012) Evaluation models and evaluation use. *Evaluation* 18(1): 61–77.

De Moor K, Berte K, de Marez L, et al. (2010) User-driven innovation? Challenges of user involvement in future technology analysis. *Science and Public Policy* 37(1): 51–61.

Dekker R, Franco Contreras JF and Meijer AJ (2019) The living lab as a methodology for public administration research: A systematic literature review of its applications in the social sciences. *International Journal of Public Administration* 43(14): 1207–17.

Engbersen G, Jennissen R and Bokhorst M (2015) *Geen tijd te verliezen: Van opvang naar integratie van asielmigranten*. WRR Policy Brief no. 4. The Hague: WRR.

European Commission (2016) Urban innovative actions. Available at: https://ec.europa.eu/regional_policy/sources/activity/urban/urban_innovative_actions.pdf (accessed 22 September 2020).

Følstad A (2008) Living labs for innovation and development of information and communication technology: A literature review. *Electronic Journal of Organizational Virtualness* 10: 99–131.

Gascó M (2017) Living labs: Implementing open innovation in the public sector. *Government Information Quarterly* 34(1): 90–8.

Geuijen K, Moore M, Cederquist A, et al. (2017) Creating public value in global wicked problems. *Public Management Review* 19(5): 621–39.

Karch A, Nicholson-Crotty SC, Woods ND, et al. (2016) Policy diffusion and the pro-innovation bias. *Political Research Quarterly* 69(1): 83–95.

Kimbell L and Bailey J (2017) Prototyping and the new spirit of policy-making. *CoDesign* 13(3): 214–26.

Kolko J (2010) Abductive thinking and sensemaking: The drivers of design synthesis. *Design Issues* 26(1): 15–28.

Lawrence RB, Rallis SF, Davis LC, et al. (2018) Developmental evaluation: Bridging the gaps between proposal, program, and practice. *Evaluation* 24(1): 69–83.

Leeuw FL and Donaldson SI (2015) Theory in evaluation: Reducing confusion and encouraging debate. *Evaluation* 21(4): 467–80.

Lewis JM, McGann M and Blomkamp E (2020) When design meets power: Design thinking, public sector innovation and the politics of policymaking. *Policy & Politics* 48(1): 111–30.

McGann M, Blomkamp E and Lewis JM (2018) The rise of public sector innovation labs: Experiments in design thinking for policy. *Policy Sciences* 51(3): 249–67.

Mackenzie M and Blamey A (2005) The practice and the theory: Lessons from the application of a theories of change approach. *Evaluation* 11(2): 151–68.

Milley P, Szijarto B, Svensson K, et al. (2018) The evaluation of social innovation: A review and integration of the current empirical knowledge base. *Evaluation* 24(2): 237–58.

Noordegraaf M, Douglas S, Geuijen K, et al. (2019) Weaknesses of wickedness: A critical perspective on wickedness theory. *Policy and Society* 38(2): 278–97.

Patton M (2011) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford Press.

Rey L, Tremblay MC and Brousselle A (2014) Managing tensions between evaluation and research: Illustrative cases of developmental evaluation in the context of research. *American Journal of Evaluation* 35(1): 45–60.

Shadish WR, Cook TD and Campbell DT (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Stoker G and John P (2009) Design experiments: Engaging policy makers in the search for evidence about what works. *Political Studies* 57(2): 356–73.

Tõnurist P, Kattel R and Lember V (2017) Innovation labs in the public sector: What they are and what they do? *Public Management Review* 19(10): 1455–79.

Verweij M and Thompson M (2006) *Clumsy Solutions for a Complex World: Governance, Politics and Plural Perceptions*. Basingstoke: Palgrave Macmillan.

Voorberg WH, Bekkers VJJM and Tummers LG (2015) A systematic review of co-creation and co-production: Embarking on the social innovation journey. *Public Management Review* 17(9): 1333–57.

Weiss CH (1997) Theory-based evaluation: Past, present, and future. *New Direction for Evaluation* 76: 41–54.

Williams M (2020) External validity and policy adaptation: From impact evaluation to policy design. *The World Bank Research Observer* 35(2): 158–91.

Rianne Dekker is an Assistant Professor at the Utrecht University School of Governance (USG). Her research interests include the governance of migrant integration and developing research methodologies to study this issue.

Karin Geuijen is an Assistant Professor at the Utrecht University School of Governance (USG). Her research interests focus on multi-level and multi-sector governance, specifically in the domain of forced migration.

Caroline Oliver is an Associate Professor of Sociology at University College London. Her research interests are in migration, state policies and institutions (welfare, education, etc.) and their consequences for social justice.