Ringa Raudla, Külli Sarapuu, Egert Juuse, Nastassia Harbuzova, Kerli Onno, Johanna Vallistu, Aleksandrs Cepilovs.
2023 "To Experiment or not to Experiment in Tax Policy?"
Halduskultuur. The Estonian Journal of Administrative Culture and Digital Governance 23(1), 27-48

To Experiment or not to Experiment in Tax Policy?

RINGA RAUDLA, KÜLLI SARAPUU, EGERT JUUSE, NASTASSIA HARBUZOVA, KERLI ONNO, JOHANNA VALLISTU, AND ALEKSANDRS CEPILOVS
RAGNAR NURKSE DEPARTMENT OF INNOVATION AND GOVERNANCE, TALLINN UNIVERSITY OF TECHNOLOGY, ESTONIA

Abstract

Despite the increasing use of experiments in policy-making the suitability of field experiments in the public sector context is still under debate. In this article, we focus on experimenting in the field of tax policy and ask: what are the promises and pitfalls of using experimental approaches in tax policy? While the existing discussions on tax policy experimentation focus on randomized controlled trials from a legal perspective, we adopt a broader view and provide a more comprehensive discussion by synthesizing insights from the fields of political science, public policy, public administration, and governance. Our analysis encompasses randomized controlled trials, non-randomized policy pilots and design experiments. We summarize the existing knowledge on using field experiments in policy-making and discuss the implications of the knowledge for experimenting in tax policy. We seek to offer a more holistic and critical take on whether we should promote the use of experimental approaches in this domain.

Keywords: policy experiments, tax policy, randomized controlled trials, policy pilots, design experiments

1. Introduction

Converging trajectories in various disciplines point to the increasing importance of using experimental approaches in policy-making (e.g. Ansell and Bartenberger 2016; John 2014; Ettelt et al. 2015a; Lee and Ma 2020; McFadgen and Huitema 2017; McGann et al. 2018). This has been aided by parallel developments in academia and practice. First, the term "experimentalist governance" has been used to describe developments in the European Union's approach to policy-making in various policy sectors (Rangoni and Zeitlin 2021). Second, the rise of the so-called "randomistas", who advocate the use of randomized controlled trials (RCTs) in development (as evidenced by Banerjee, Duflo and Kremer getting the Nobel prize in economics in 2019), has strongly influenced debates on development economics, with spillovers to other areas as well (e.g. Kvangraven 2020; Leao and Eyal 2019; Stein et al. 2021). Third, experimentalism has been embraced by behavioral economics and behavioral public policy (e.g. Jones and Whitehead 2018; Strassheim 2020). Fourth, the calls to integrate design thinking – which is also experimentalist in its logic – into policy-making have recently been made by a number of public policy scholars (e.g. Stoker and John 2009; McGann et al. 2018). In response to the pressures to be "smarter" and more "innovative", many governments have created nudge-type government units and public sector innovation labs that employ various

kinds of experimental methods (e.g. John 2014; McGann et al. 2018; Tõnurist et al. 2017).

At the same time, the world is becoming increasingly uncertain and complex, warranting experimentation as a way to cope (Rangoni and Zeitlin 2021; Voß and Simons 2018). For example, the need to deal with climate change has precipitated an avalanche of sustainability experiments (e.g. Ansell and Bartenberger 2016; McFadgen and Huitema 2017, 2018 McFadgen 2019). Fast technological developments (especially in ICT) and their pervasive impacts on economic and social spheres have triggered extensive discussions on experimental legislation and regulation (e.g. Philipsen et al. 2021; Ranchordas 2013).

The suitability of experimenting in the public sector context is still under debate, especially with regard to specific policy fields. In this article, we focus on experimenting in the field of tax policy. More specifically, our research question is: What are the promises and pitfalls of using experimental approaches in tax policy? The focus has been inspired by various considerations. Tax policy entails the use of taxes for a number of purposes: generating revenue for the government, shaping the behavior of individuals, redistributing income, and stabilizing the economic cycle (e.g. Kay 1990). Given that *consistency* and *horizontal equity*[1] are considered to be the core values in this policy area, using experimental approaches – which by nature entail disruption and potentially differential treatment of citizens – is likely to be more controversial in tax policy than in many other policy fields. At the same time, it is a policy sector where the complex and rapidly evolving context (e.g. changing technology, globalization, challenges presented by climate change) gives rise to considerable new uncertainties. Experiments could be viewed as generating useful knowledge about the novel challenges and the best ways to address them (Rangoni and Zeitlin 2021). Given the importance of taxes in generating revenues for the state, getting tax reforms "right" has very high stakes (Werner and Riedl 2019).

Recently, rather bold proposals for making more extensive use of field experiments in tax policy have been put forward by Abramowicz (2019). While Abramowicz (2019) approached the issue from the legal perspective and focused specifically on the promises of *randomized* experiments, we adopt a broader view. We provide a comprehensive discussion about the suitability of policy experiments in the field of tax policy by synthesizing insights from the fields of political science, public policy, public administration, and governance. With regard to the different *types* of experiments, our analysis encompasses both randomized controlled trials and non-randomized policy pilots as well as design experiments. We summarize the existing knowledge on using experimental approaches and discuss the implications of the knowledge for experimenting in the field of tax policy. With that, we hope to offer a more holistic (and also a more critical) take on whether we should promote the use of experimental approaches in this domain.

It is worth emphasizing that in this article we focus on tax *policy* rather than tax *administration*. The use of field experiments to improve tax compliance (e.g. by using deterrence messages, referring to the tax behavior of others, increasing the moral costs of non-compliance in tax offices' communication with taxpayers) has been extensively discussed in the existing studies (for useful overviews, see Hallsworth 2014; Mascagni 2018; Pomeranz and Vila-Belda 2019). Furthermore, our focus is not on whether *research* in public administration, public policy and political science should make (more) use of experimental methods (as discussed e.g. in Druckman et al. 2006; Stoker 2010). Instead, we are interested in whether *policy-makers*, in

1 Consistency refers to taxing equivalent transactions in a similar way. Horizontal equity means that taxpayers who are in a similar position should face a similar tax burden (e.g. Elkins 2006; Musgrave 1990).

devising and improving policy measures in the tax policy domain, should make more extensive use of experimental approaches, as the current *Zeitgeist* seems to be prescribing. Also, our focus is on field experiments and not on laboratory experiments.[2]

The article is structured as follows. Section 2 explains what we mean by the term "policy experiment". Sections 3 and 4 give an overview of the potential benefits and pitfalls of making more extensive use of experimental approaches in tax policy, respectively. While sections 3 and 4 focus on issues that are relevant for all experimental approaches (irrespective of the specific design choices), section 5 analyzes the specific promises and challenges presented by different types of designs. Section 6 concludes.

2. Definition of policy experiment

Policy experiment is a protean concept (Karvonen and van Heur 2014), and in the existing literature, various definitions have been offered (for overviews of these debates, see, e.g., Ansell and Bartenberger 2016; Bauknecht et al. 2020). In some disciplines (e.g. economics), the term "experiment" typically has a rather narrow meaning and refers to a randomized controlled trial (RCT) (Ansell and Bartenberger 2016; Burtless 1995; Werner and Riedl 2019). On the other extreme, any policy reform – like the reduction in the top marginal rates of the income tax in the US in the 1980s – has been labelled as an "experiment" (Burtless 1995; Druckman et al. 2006; Heldeweg 2015; Nair and Howlett 2016).

In this article, we opt for the middle ground. The definition of policy experiment we employ is as follows: it is a policy relevant test undertaken by government organization(s) to learn about the impacts of a new policy solution, which can be used as evidence for further policy decisions (Bravo-Biosca 2020, 195; Heldeweg 2015, 183; McFadgen and Huitema 2017, 1768; Nair and Howlett 2016, 69; Millo and Lezaun 2006, 179). In an experiment, a new policy solution is "tried out or tested in a restricted environment in terms of time, space, scope and/or actors", but it is "intended to provide a proof of principle that subsequently could have the potential of wider societal relevance" (Heldeweg 2015, 183). Experiments are "temporary and reversible interventions without permanent policy consequences" (McFadgen and Huitema 2017, 1767).

We regard a policy experiment as "a process that generates learning through an explicit intention to test new ideas" (McFadgen and Huitema 2017, 1765). As Bravo-Biosca (2020, 195) explains, an experiment is "intentionally set up to learn", has "a clearly structured learning strategy", and generates "new information, evidence, or data". Therefore, if a government just tries out something new, it does not amount to a policy experiment unless the systems and processes required to learn from it are also established. This includes a timeframe for checking results and deciding whether to continue the experiment, adjust it, discontinue or scale up (Bravo-Biosca 2020). In other words, we follow the understanding whereby policy experimentation does not entail "freewheeling trial and error or spontaneous policy diffusion" but is "purposeful and coordinated activity geared to producing novel policy options that are injected into official policymaking", with the goal to scale them up, if successful (Heilmann 2008, 3).

2 Useful overviews of laboratory experiments that could be insightful for tax policy have been provided by Werner and Riedl (2019) and Alm and Malézieux (2021).

3. The promises and benefits of using experimental approaches

The advocates of the experimental turn posit that experimental approaches can improve the substantive quality of policies as well as the policy-making process. First, from the perspective of *policy content*, the key aspects that speak for experimentation are the complexity of the system policy-makers seek to influence and uncertainty about the impacts of an intervention (Ansell and Bartenberger 2016; Bauknecht et al. 2020; Bravo-Biosca 2020; Hughes et al. 2020; Lee and Ma 2020; Lee et al. 2009; Millo and Lezaun 2006; Nair and Howlett 2016; Voß and Simons 2018). Experimental approaches are considered especially useful for understanding complex systems where actors, institutions and policy continuously evolve and interact in various ways (Bravo-Biosca 2020) or are influenced by novel and disruptive technological developments (Bauknecht et al. 2020; Philipsen et al. 2021; Ranchordas 2013, 2015a, 2015b; Van Gestel and Van Dijck 2011).

Therefore, the key promise of policy experiments, regardless of the specific design, is to provide policy-makers with information that would otherwise not be available (Abramowicz 2008; Ettelt et al. 2015b; Millo and Lezaun 2006; Philipsen et al. 2021). Experiments allow policy-makers to test the effectiveness of a policy on a smaller scale, but in a real world setting, before it is rolled out on a larger scale (Bailey et al. 2017; Checkland et al. 2021; Farrelly 2008; Hughes et al. 2020; Lee et al. 2009; Philipsen et al. 2021; Ranchordas 2013). In addition to allowing the exploration of whether intended impacts materialize, experiments can offer information about unintended and adverse effects (Haynes et al. 2012; Hughes et al. 2020; Lee and Ma 2020; Nair and Howlett 2016; Philipsen et al. 2021; Ranchordas 2013), reduce informational asymmetries with regard to policy's acceptability for stakeholders (Nair and Howlett 2016 Philipsen et al. 2021), and alleviate uncertainty about how the target group behaves in response to different measures (Burtless 1995; Nair and Howlett 2016; Ranchordas 2013). Similarly, experiments can offer information regarding the impacts of different policy design options and help policy-makers choose between them (Bravo-Biosca 2020; Haynes et al. 2012; Hughes et al. 2020; Lee and Ma 2020; Nair and Howlett 2016). In the context of disruptive technological developments, experiments help to diminish the pacing and information gaps between technological innovations and legislation (Ranchordas 2013, 2015c).

These promises of experimenting reflect the information needs also in the field of tax policy. For example, if policy-makers were considering the use of tax credits for promoting the adoption of more environmentally sustainable investments and business practices by companies, an experiment could offer information about the degree of take-up of such incentives. Since taxpayers' perceptions of tax incentives can be distorted (Werner and Riedl 2019), it can be difficult to make linear predictions about their effects. Experiments can offer useful insights into how such incentives are actually perceived. With regard to choosing the best tax policy design, an experiment may allow policy-makers to compare the effects of tax credits, tax deductions and tax exemptions, and opt for the solution that is likely to have the largest impact on the adoption of environmentally sustainable practices. In assessing the effects of tax credits on the behavior of businesses, policy experimenters can examine whether the size of the tax credit makes a difference. Experiments could also be used for assessing which aspects of the policy are having the largest effect (Haynes et al. 2012). For example, if the government employs, in parallel, tax credits but also subsidies for promoting environmentally

sustainable investments, it may be interested in which instrument is responsible for driving the policy outcomes. In an era of disruptive technological developments, experiments can be useful for testing how to best tax digital nomads, digital transactions, digital data, carbon, or ecological footprints (Akdogan 2021; Shome 2021). In such instances, experiments also allow governments to test their own ability to impose taxes on novel phenomena (Millo and Lezaun 2006).

Second, from the perspective of *policy process*, experimentation can improve the quality of deliberations over policy. In uncertain situations, policy experimentation can help to build consensus and create space for political bargains by focusing policy-makers' attention on policy consequences and outcomes (Abramowicz 2008; McFadgen 2019; Millo and Lezaun 2006; Nair and Howlett 2016; Lee and Ma 2020; Ranchordas 2013, 2015a, 2015c). This can be particularly beneficial in tax policy, where the partisan debates tend to be polarized and more influenced by policy hunches (or even dogmas) and less by information about the actual impacts of different solutions (Abramowicz 2019). By presenting a possible future, experimenters can mitigate conflict in policy reforms that may otherwise be politically unpalatable (Bailey et al. 2017; Nair and Howlett 2016; Ranchordas 2013, 2015a, 2015c). Consensus on a policy, in turn, can potentially facilitate higher levels of compliance and, through that, increase policy effectiveness.

Furthermore, experimentation can be viewed as helping to lower the overall costs of policy (Abramowicz et al. 2010; Adkins and Ylöstalo 2018; Checkland et al. 2021; Farrelly 2008; Haynes et al. 2012; Hughes et al. 2020; Kvangraven 2020; Lee et al. 2009). Even though the running of experiments can entail data collection costs, such investments can pay off via allowing policy-makers to "weed out" programs that are not effective or avoid the costs of failure (Abramowicz et al. 2010; Bravo-Biosca 2020; Farrelly 2008; Haynes et al. 2012; Oakley 1998; Lee and Ma 2020; Lee et al. 2009). Experiments can also help to decide, in the context of limited resources, which policy options from a range of alternatives deliver the highest value and should be chosen over the others (Haynes et al. 2012). Given that changes in tax policy can entail considerable costs for the state budget (especially when they entail tax incentives), the financial considerations in opting for experiments can be particularly pertinent in the field.

4. The challenges and constraints of using experimental approaches

The more critical perspectives on using experiments in public policy point out several challenges and constraints. These derive from two sources – the problems inherent to experimental approaches and the political setting of public policies. Tax policy experiments are likely to be challenged by both.

There are challenges that are common to all types of policy experimentation, regardless of their specific design – whether they "slice" through space, time, scope or types of participants. For instance, an important challenge for all experiments is that due to their small scale, scope or limited duration they would not be able to capture some of the effects which could materialize when the same policy is offered at a large scale or over a longer period of time or in a different context (Bauknecht et al. 2020; Burtless 1995; Millo and Lezaun 2006; Werner and Riedl 2019). As Millo and Lezaun (2006, 181) put it, one can always point to "particular features of the world that the experiment failed to replicate". For example, in an experiment

that tests the impacts of tax credits for investments to lower the carbon footprint, if only a small fraction of businesses is enrolled, it may not be able to capture the kinds of peer-effects and isomorphic pressures that may emerge if the scheme applied to all businesses. Thus, in tax policy experiments, one should remain aware of the dangers of making "sweeping macro-level generalizations" by drawing on micro-level evidence (Stein et al. 2021, 64). Attribution of causality can be particularly challenging when "complex systems are involved in a difficult task" (Nair and Howlett 2016, 71). Thus, tax policy experiments would face challenges in assessing third-party effects (i.e. effects that go beyond the taxpayers themselves) and especially macroeconomic consequences of tax changes (Abramowicz 2019).

Furthermore, a key challenge for all experiments, including those in tax policy, is the Hawthorne effect, which occurs when individuals or businesses behave differently because they know that they are in an experiment (e.g. Levitt and List 2011). Also, if the participants know that the treatment is of limited duration, their reactions may be different from how they would react if the same policy was of enduring character (Burtless 1995). For example, taxpayers subject to an experiment may try to lower their tax payments by shifting their income and deductions from years in which they are subject to the experiment to years when the generally applicable tax rules apply to them or the other way around (Abramowicz 2019; Abramowicz et al. 2010). In such cases, a possible solution could be to have a longer time period for the experiment (spanning several years) (Abramowicz et al. 2010), which would limit the ability of individuals or companies to shift their income or expenses. However, the longer the experiment, the more serious the problem of attrition (Burtless 1995), which challenges the assessment of impacts. In addition to the Hawthorne effect, the results of an experiment may be influenced by spillovers to areas or individuals who were not supposed to be affected by the experiment. For example, if tax incentives for individuals or companies are offered in one jurisdiction (e.g. local government) and not in others, taxpayers may relocate to the experimental jurisdiction, which makes it difficult to compare the experimental jurisdiction with the non-experimental ones (Abramowicz 2008).

Next to the constraints related to the general nature of experiments, the *political setting* of public policies in a democratic context poses an additional range of challenges. The prevailing logics of political decision-making, policy and electoral cycles, electoral considerations, and value conflicts can render experimental approaches difficult for policy-makers to undertake (Bauknecht et al. 2020; Bravo-Biosca 2020; Burtless 1995; Voß and Simons 2018). The political setting entails four types of causes that contradict experimental logic and, consequently, may inhibit undertaking the experiments or shape their implementation.

The first challenge is related to the nature of politics per se. Politicians may reject policy experimentation as a general strategy if they regard it as a route for depoliticizing or de-democratizing policy decisions (Pearce and Raman 2014; Voß and Simons 2018; Strassheim 2020). Extensive literature about the obstacles to evidence-based policy-making has suggested that politicians often follow other logics than systematic evidence in adopting policy decisions. In a democratic system, competing societal values and preferences are translated into policy solutions (Kvangraven 2020). Thus, politicians may attribute higher importance to voter preferences, party agendas, and ideological considerations rather than to the kind of evidence that experiments promise to deliver. This is especially relevant for tax policy. Tax issues often constitute salient agenda points in electoral competition (Klitgaard

et al. 2015) and can define the core identity of some parties (Carmines et al. 2012; Ballard-Rosa et al. 2017; Osterloh and Debus 2012). As a result, tax policy issues may be regarded very ideologically and viewed as the core space in which ideological trade-offs (e.g. more state vs less state, equity vs freedom etc.) are struck (Carmines et al. 2012). This makes it particularly challenging to use experimental approaches in this domain. Driven by ideology, politicians may trust their gut feeling about the potential impacts of different tax policies – e.g. higher rates are "good" or "bad" and exemptions are "good" or "bad" depending on the ideological spectrum they are on – and hence do not feel the need to demand more rigorous evidence via experiments (Bravo-Biosca 2020). Politicians are likely to object to any tax experiments that they perceive to hurt their re-election chances (Bauknecht et al. 2020; Nair and Howlett 2016). Conversely, a tax policy option may be "so in tune with prevailing political values that subjecting it to tests is regarded by politicians as unnecessary or even unwelcome" (Pearce and Raman 2014, 393).

The second issue pertains to the need to avoid uncertainty. Beyond any ideological considerations, politicians may be concerned about how experimental decision-making as a policy-making style looks to the electorate. Experimentation may be problematic for politicians, since undertaking an experiment entails admitting uncertainty and lack of existing knowledge about the effects of policy interventions (Bauknecht et al. 2020; Farrelly 2008). Thus, if politicians were to opt for experimental tax policy at all, they may prefer to treat experiments as "demonstrations of effectiveness" or "justifications for decisions already taken" rather than answering open questions (Ettelt et al. 2015a, 294, 302; see also Zurbriggen and Lago 2019). These concerns may be reinforced by the culture of public sector organizations, which tend to be risk-averse and prefer stability, predictability and order, and exhibit limited tolerance of failure (Lewis et al. 2020; McGann et al. 2018; Zurbriggen and Lago 2019). The need to avoid uncertainty is likely to be particularly pronounced in case of a core state function like taxation. Instead of admitting uncertainty, elected representatives may prefer to be perceived as "decisive, energetic, and positive" (Farrelly 2008, 11) and emphasize "profound knowledge" about future developments (Bauknecht et al. 2020, 57). Experimenting strikes the kind of "chord of skepticism and indecision" that politicians seek to avoid (Peters 1998, 126) – even if the policy change in question entails using tax policy instruments in new technological spheres (e.g. taxing digital transactions) or changing societal context (e.g. digital nomads or platform work), all facing considerable uncertainties. Although the higher the uncertainty, the more learning opportunities experiments can offer, politicians may regard an open outcome as entailing a higher likelihood of failure – and this may be something they are motivated to avoid (Ettelt et al. 2015b; McFadgen and Huitema 2018).

The third issue concerns constrained timeframes. Some policy decisions may require immediate decisions, and hence policy-makers may not have the time needed for experimenting (Clarke and Craft 2019). For example, when tax policy is used for macroeconomic stabilization (i.e. cutting rates during recession and increasing them during a boom), timing is of pivotal importance and swift decisions crucial if undesirable lagged effects on the economic cycle are to be avoided. Furthermore, politicians may perceive that they are likely to be electorally rewarded for being swift rather than slow. As Stoker (2010, 53) explains, the demands of the experiment may clash with the policy cycles or political dynamics: "Experiments are a tool with a linear rhythm in a non-linear policy process and may as a result lose the battle for relevance by failing to produce results in a timely way." There might be political pressure

to go fully ahead with the policy owing to its perceived benefits rather than waiting for the results of the experiment (Stoker 2010) or hasten the evaluation of the pilot, especially if the policy deals with pressing social problems (Nair and Howlett 2016). Especially in the case of longer-lasting experiments spanning several years (like the social experiments in the US in the 1960s and 1970s), the political agendas may change, rendering the findings less relevant for policymakers (Burtless 1995; Oakley 1998).

Fourth, there may be lacking capacities. Experimental policy-making requires various types of analytical and collaborative capabilities from the policy-makers, which may be lacking (Bedard and Ouimet 2012; Bravo-Biosca 2020; Stoker 2010). Lacking collaborative capabilities can become a crucial hurdle to policy experiments that involve a large number of different organizations and the engagement of public officials at several levels of government (Stoker 2010, 51; Cotterill and Richardson 2010, 157). While some simpler tax policy experiments may be relatively straightforward and just involve the finance ministry and tax office, others (e.g. on using tax incentives for fostering innovation or environmental sustainability) can involve a considerably larger number of bodies. Furthermore, objections to experiments may emerge from the broader public, and considerable communicative efforts may be needed to explain the social benefits of an experiment to gain public acceptance (Bauknecht et al. 2020) – especially in domains like tax policy where the public is likely to have deep-rooted expectations of horizontal equity.

5. Promises and pitfalls of specific experimental designs

Policy experiments can take different forms. In this section, we will focus on the following designs: randomized controlled trials, non-randomized policy pilots, and design experiments. These are the main "ideal types" of experiments that have been discussed in the experimental policy-making literature. They follow different logics and hence entail different benefits and challenges from the perspective of tax policy. While sections 3 and 4 outlined the benefits and challenges that are common to *all* experimental approaches in tax policy, in this section, we zoom in on the promises and pitfalls of these three specific designs.

5.1. Randomized controlled trials

Randomized controlled trials (RCT) hold the promise of allowing the experimenters to draw valid *causal conclusions* about the effects of a project, program or policy (Bravo-Biosca 2020; Burtless 1995; Cook 2002; Dalziel 2018; Ettelt and Mays 2015; Haynes et al. 2012; Pearce and Raman 2014). The random division of subjects into experimental and control groups can be expected to eliminate systematic differences between them and create equivalent groups, which are then subjected to different treatments (in the simplest design, intervention for the experimental group and nothing for the control group) (Burtless 1995; Cook 2002; Ettelt and Mays 2015; Farrelly 2008; Pearce and Raman 2012). In such a design, any observed differences between the groups are attributed to the "treatment" (i.e. the tested policy, program or project), assuming that the experimental group and control group operate in the same policy, social and economic environment (Bell and Peck 2016; Cook 2002). An RCT is expected to create a credible counterfactual (in the form of a control group), and this should enable

policy actors to assess the average treatment effect of the intervention (Bedard and Ouimet 2012; Bravo-Biosca 2020; Burtless 1995; Farrelly 2008; Haynes et al. 2012).

An RCT is widely regarded as "a gold standard method for measuring whether or not a particular intervention works better than doing something else or doing nothing" (Cotterill and Richardson 2010, 156). Given the uncertainties involved in many tax policy measures, RCTs could, in principle, offer opportunities to shed light on these questions in a systematic way.[3] An RCT could be used to assess the additional value (Bravo-Biosca 2020) generated by tax incentives, for example. It could offer the opportunity to test the assumptions or intuitions policy-makers have about a new tax deduction or exemption in terms of their behavioral effects or a new tax on previously untaxed objects or activities. As Haynes et al. (2012) emphasize, the untested intuitions of policy-makers may be wrong, even with policies that should be "guaranteed" to work.

Despite being regarded as the "gold standard" for causal inference (Bravo-Biosca 2020; Strassheim 2020; Webber and Prouse 2018), we should be aware of the dangers inherent in such "methodological triumphalism" (Barrett and Carter 2010, 516) and also pay attention to the key challenges of RCTs. RCTs suffer from major shortcomings – in light of which one could claim that, despite the glitter of recent Nobel prizes, it is "of baser metal than gold" (Barrett and Carter 2010, 516). Indeed, these may be the reasons behind the fact that although "tax law is a promising field in which the government might run randomized experiments", existing experiments only entail tax compliance and welfare (i.e. negative income tax experiments) rather than tax policy more broadly (Abramowicz 2019, 68). In the following, we will discuss the challenges of RCTs.

The RCT design works well under three assumptions: 1) the intervention has to be clearly delineated; 2) the expected outcomes of the intervention have to be measurable and identified in advance of the experiment, and 3) the causal mechanisms examined should be relatively simple (Bedecarrats et al. 2019; Jones and Whitehead 2018). Although such a setup allows for valid statistical conclusions about the average treatment effect (Dalziel 2018; de Leao and Eyal 2019; Kvangraven 2020), it also imposes considerable limitations on what kinds of policies we can test with such a design (Bedecarrats et al. 2019). All of these aspects may pose challenges for using RCTs in tax policy experimentation.

First, RCTs rely on a definition of policy as an "intervention" while actual conceptions of policy are usually more diffuse (Ettelt and Mays 2015, 380). Some aspects of tax policy – e.g. marginal tax rates – can be easily operationalized and quantified for the purposes of an RCT. For example, policy makers may test with an RCT whether lower tax rates for hiring disadvantaged workers have the intended effect. However, many other tax policy domains may be more diffuse. Tax policies, like most other policies, are often constituted by configurations of interacting activity setting, events, and technologies rather than by discrete interventions (Anderson 1975; Ettelt and Mays 2015). As Anderson (1975, 17-18) explains, even in a relatively straightforward negative income tax experiment[4], the "policy treatment" would be characterized by a host of variables such as rules for family size, what counts as a household, household benefits from other sources, accounting periods, and how windfall income is treated. All these are likely to affect the costs of the policy and its efficiency.

3 Indeed, some of the "classic" RCTs during the era of social experiments (1960-1980) in the US used the tax system as a policy instrument – most famously the experiments on negative income taxes (Burtless 1995; Oakley 1998).

4 Negative income tax means that the government sends money back to taxpayers whose income falls below a certain threshold (Moffitt 2003).

Second, RCTs are suitable when policy outcomes are easily measured but challenging when they are fuzzy (Bedecarrats et al. 2019; Bravo-Biosca 2020; de Leao and Eyal 2019). Thus, in the case of broad-aim tax policies, which entail wider societal impacts, general equilibrium effects or changes in economy-wide aggregates, the use of RCT would be difficult (Bedecarrats et al. 2019; Burtless 1995; Deaton 2010; Strassheim 2020). As Burtless (1995, 77) explains, a negative income tax experiment, for example, would be able to capture the effects on labor supply, but without knowing how the employers would alter the wages, it would be impossible to "forecast the full general equilibrium effect".

Furthermore, RCTs work well for assessing *average* effects but are challenging when the policy outcomes are skewed – e.g. when most projects fail and extreme successes are rare, as could be the case in tax-incentivized investment projects (Bravo-Biosca 2020; Bedecarrats et al. 2019; Deaton 2010). For instance, the take-up of tax credits for environmentally sustainable investments may be influenced by exceptional managerial capabilities (a characteristic that is difficult to measure), and this can introduce inaccuracies when estimating the average effect (Dalziel 2018). Similar objections are likely to be present in other tax policy measures that entail high variability in the target group. There might be considerable heterogeneity in how different groups respond to the tax policy treatment (Werner and Riedl 2019) and hence the average effect may conceal the fact that some groups are responding strongly and others are not. Furthermore, while some of the questions in tax policy may zoom in on *average* treatment effects, most of the issues policy-makers might be interested in concern *conditional* effects (e.g. the effects of tax incentives on the most innovative companies or the effects of tax credits on the poorest families). Policy-makers may also be interested in the *distribution* of positive and negative effects in different societal groups or types of firms, rather than just the aggregate average effect (Barrett and Carter 2010; Bedecarrats et al. 2019; Deaton 2010).

Third, "randomized experiments are best when a causal question is simple, sharply focused and easily justified" (Cook 2002, 179). If the policy intervention is targeted at a complicated phenomenon with a complex ecosystem (including unobserved interactions and linkages) it may be difficult to predict the impacts (Bravo-Biosca 2020; Bedecarrats et al. 2019). This is likely to be the case in experimenting with new forms of taxes on digital transactions or carbon footprints, for example. Also, those aspects of tax policy that seek to affect overall macroeconomic outcomes (e.g. stimulating the economy via lowering tax rates) depend on collective actions and interactions – and this clearly undermines the case of using RCTs for studying them.

RCTs are also challenging when the causal processes through which policies affect outcomes take *a long time* (Bravo-Biosca 2020). For reasons of cost and attrition, RCTs tend to be short in duration, which means that in reality only mid-point measurements rather than final indicators can be captured as outcomes (Bedecarrats et al. 2019; Farrelly 2008). Furthermore, the longer the time span of the experiment, the higher the likelihood that other factors besides the intervention start influencing the outcome(s) (de Leao and Eyal 2019; Farrelly 2008). For example, the effects of tax credits for investing in environmentally sustainable technologies may take a long time to materialize, and short-term evaluation may present an inaccurate picture of the eventual effects.

In addition to the aforementioned challenges, the design feature of using *randomization* can pose considerable obstacles to tax policy experiments. In the field of tax policy, randomization

may be difficult to justify ethically, politically, and legally.

Ethically, it may be difficult to argue why a policy measure should be denied to some potential recipients who would benefit from it (Bravo-Biosca 2020; Burtless 1995; Jones and Whitehead 2018; Pearce and Raman 2014). For example, if some companies have the opportunity to make use of tax credits for environmentally sustainable investments and others do not, this may be seen as unfair. Furthermore, randomization inherent in RCTs is especially likely to give rise to ethical objections when the experiment involves target groups that are vulnerable (Cotterill and Richardson 2010; Cook 2002). This may make it challenging, for example, to use an RCT to test the effects of giving tax incentives to employers for hiring disadvantaged workers.[5]

Politically, randomization may be difficult to justify to the wider public (Nair and Howlett 2016; Strassheim 2020). Policymakers may be concerned about whether it is fair for some people to receive help or benefits and others not if the experiment uses public resources (Nair and Howlett 2016; Strassheim 2020). It may be even more difficult for politicians to impose additional tax burdens on some individuals and businesses but not on others. For example, if the government wanted to try out a car tax on a smaller scale (e.g. applying it to randomly selected car owners) before implementing it on a large scale, in order to examine how it influences people's consumption choices, it may be next to impossible to justify it politically. Similarly, trying out the effects of a carbon tax (e.g. on the investment capacity of firms) in certain parts of the country before extending it to the whole country may run into similar difficulties. If randomization is difficult to justify, policy-makers are likely to fear negative public backlash to the trials (Bravo-Biosca 2020; Strassheim 2020) – and in the domain of tax policy, with clearly measurable costs and benefits, voters might be perceived to be particularly sensitive with regard to being treated unfairly.

Legally, challenges may arise from treating people differently when they should be treated equally (Adkins and Ylöstalo 2018; Burtless 1995; de Leao and Eyal 2019), especially if the intervention concerns rights or obligations of citizens (Abramowicz et al. 2010; Burtless 1995). In tax law, as Abramowicz (2019, 69) argues, the main hurdle to RCTs is the "core value of horizontal equity", which leads to "concerns that experiments necessarily produce unequal treatment of similarly situated individuals." However, Abramowicz (2019) argues that there might still be ample room for tax policy experiments where randomization could be legally justified with the fact that the experiment is revenue neutral. Revenue neutrality means that the "treatment group in such an experiment on average pays taxes as high as the control group" (70). In particular, such revenue-neutral designs would be suitable for assessing which combinations of tax deductions and tax rates would be most efficient (from the point of view of allocative efficiency). For example, if the current tax system entails specific deductions but higher (marginal) rates, policy-makers may want to test, with an RCT, whether taxpayers could in fact be better off with abolished (or reduced) deductions but lower rates. For instance, a government could undertake an experiment on abolishing entertainment deductions for businesses: businesses participating in the experiment would give up these deductions but face lower overall tax rates. Conversely, if the current system offers no deductions but lower marginal tax rates, policy-makers may want to experiment with offering deductions, combined

5 A solution proposed to make randomization more palatable to the public is in framing the experiment as "a lottery". For example, in an RCT that has the potential to yield valuable insights in the effects of different marginal tax rates on incentives to work or for entrepreneurship, it may be very difficult to justify to the public why some households are subjected to higher tax rates than others. Abramowicz et al. (2010, 999) suggest a solution according to which the government announces that it "is sponsoring a lottery, the winners of which receive a reduction in their tax rates".

with higher tax rates. Abramowicz (2019) suggests that in addition to small changes to tax systems (e.g. the effects of adding or removing deductions), revenue neutral randomized experiments could be used to test also more significant changes – for example replacing corporate income taxes with government equity in corporations' stock.

5.2. Non-randomized policy pilots

Given the numerous challenges involved in RCTs, they may often not be feasible options for experimental tax policy. In that light, non-randomized policy pilots may offer an alternative route. A policy pilot seeks to test out a new policy approach in a confined setting, and/or on a small subset of the population or jurisdictions, and allows the introduction of a policy in a phase-wise manner (Bailey et al. 2017; Farrelly 2008; Ko and Shin 2017; Nair and Howlett 2016). Unlike RCTs, they do *not* entail randomization, although they may include comparisons with a control group (Philipsen et al. 2021). In order to be viewed as experiments, however, pilots should entail the establishment of concrete systems or processes to learn from them (e.g. Ko and Shin 2017; Lee et al. 2009; Philipsen et al. 2021).

Pilots enable policymakers to assess and adjust a solution before rolling it out nationally (Ko and Shin 2017; Nair and Howlett 2016). For example, a government may be interested in whether tax exemptions applied to cooperation agreements between universities and companies (e.g. creation of industrial professorships) or reducing the payroll tax of high-level researchers working at companies could facilitate the commercialization of basic research. Such exemptions may be first tested with some universities and companies before applying them to all. Special tax regimes for inbound workers could first be offered in some regions of the country to test whether this could be a useful instrument for regional development. In the EU context, tax policy pilots can also take the form of testing different tax solutions in the member states.[6]

Furthermore, owing to their small scale, pilots can foster policy innovations and aid the development of new policy designs (Nair and Howlett 2016). Policy pilots can also facilitate learning how to overcome implementation barriers and improve processes (Ettelt et al. 2015b; Ko and Shin 2017), which might be crucial in the case of taxing novel phenomena. In addition, pilots can catalyze the adoption of policy innovation through *demonstrating* how a new policy can be implemented successfully (Checkland et al. 2021; Ettelt et al. 2015b; Hughes et al. 2020).

Compared to RCTs, which require clearly defined input-output-mechanisms, non-randomized tax policy pilots may be more flexible in their setup and allow a more holistic assessment of a new measure. Also, RCTs tend to take longer time and hence policy-makers may prefer to undertake simpler policy pilots in order to get the information faster (and in line with the electoral cycles) (Ko and Shin 2017), which may be an important advantage in tax policy. There are, however, a number of challenges faced by non-randomized policy pilots.

First, in contrast to RCTs, the causal claims would be significantly weaker since we do not know "how the targeted population would have fared in the absence of treatment" (Farrelly 2008,

6 For example, Directive 1999/85/EC foresaw the possibility for member states to apply a reduced VAT rate on labor-intensive services to create jobs. The evaluation of the experiment by the Commission in 2003 revealed that the reduced tax rates were not translated into lower consumer prices and no clear impact on employment rates could be identified. Thus, it was concluded that "such measures were usually not very effective and the cost to the budget was high in relation to any impact the measures might have on the economy" (Van Gestel and Van Dijck 2011, 545).

8). Effects can be difficult to assess because "there may be alternative explanations for any observed changes" (Abramowicz 2008, 34). This means that in a politically contested policy domain like taxation, the findings of non-randomized pilots can be more easily attacked by political actors who are not pleased by the results of the pilot. Since there is no randomized control, policy makers cannot be confident whether in a policy pilot tax revenues have fallen because of the piloted measure or due to exogenous reasons, such as economic recession (Abramowicz 2019).

Second, replicating the success of a policy pilot may be challenged by differences in context (van der Heijden 2018; Farrelly 2008). The findings of policy pilots are likely to be influenced by various interdependent social, political, and economic factors, and this limits external validity. The groups, organizations or regions analyzed in a pilot may be systematically different from the rest of the population (Bailey et al. 2017). Furthermore, participants in the pilot may be motivated to ensure success and have incentives that may not be shared by the broader population (van der Heijden 2018). Van der Heijden (2018, 1385) refers to a frontrunner paradox, as experiments often "look for actors who want to be actively involved in solving a problem, who do not mind deviating from routines, and who are willing to take risks." These attitudes may not be characteristic to a broader population or a larger set of organizations, making scaling up challenging. For example, regions involved in a pilot testing a new tax regime for inbound workers may be particularly motivated to use that measure for promoting economic development and engage in extensive communicative efforts in spreading knowledge about that option.

Third, due to the limited geographical scope of policy pilots, spillovers to or from other regions may constitute an important challenge, especially in tax policy, where the incentives to move across jurisdictional boundaries are likely to be significant. As economic activity can move across jurisdictions, firms may shift its activities from a more highly taxed pilot location to one with a lower tax or the other way around (Abramowicz 2019), posing considerable challenges in making accurate assessments of the actual effects of the pilot.

Fourth, similarly to RCTs, an important question tax policy pilots have to wrestle with is the question of meaningful duration. If the designated time period is too short, we may not be able to capture the full impacts of the tested policy. Ranchordas (2015a, 912) suggests that this can be pertinent in experimental tax legislation that seeks to stimulate investment in renewable energy and advance clean-energy innovation. Since it takes a long time to develop a wind farm, for example, uncertainty with regard to the renewal of such tax credits can undermine long-term investment.

Finally, analogously to RCTs, tax policy pilots may give rise to ethical, legal and political challenges (Bauknecht 2019; Van Gestel and Van Dijck 2011). For example, offering a special tax regime in some geographical jurisdictions or to some organizations may be seen as violating the principles of horizontal equity, equal treatment and legal certainty (Huitema et al. 2018; Philipsen et al. 2021). However, it may be more feasible to use objectively defined and politically justifiable criteria for picking the subjects in a tax experiment in a non-randomized rather than a randomized way. For example, offering preferential tax regimes first to the least developed geographic regions or offering tax exemptions to the most promising cooperation networks of universities and businesses may be more palatable to politicians and the electorate than randomization.

5.3. Design experiments

In a design experiment, "a solution concept (an idea, design, program, project, and so on) to a particular problem is created, and iteratively refined based on continuous feedback from stakeholders immersed in the experiment" (Zurbriggen and Lago 2019, 440). Design experiments draw on design thinking that emphasizes the importance of systems thinking, user centrism, regular iteration, and creativity (Clarke and Craft 2019, 6). Such an experiment adopts a "probe and learn" strategy, in order to understand the intervention (Ansell and Bartenberger 2016, 68) and to "re-specify and re-calibrate" the solution until it works (Stoker and John 2009, 358). Thus, the experiment would progress through iterative cycles of design, real-world testing and redesign based on lessons from earlier iterations (Stoker and John 2009, 256; van der Heijden and Hong 2021, 1119). Design experiments are less concerned with exploring causality than with manipulating an intervention in order to reach an acceptable outcome (Stoker and John 2009). Unlike in RCTs where policy experimenters can remain detached from the context, design experiments entail immersion in "thickly experiential policy contexts" (Lewis et al. 2020, 116).

A key feature of design experiments is taking into account the experiences of the persons affected by the issue. Design experiments assume that in order to address societal (especially wicked) problems, expertise from both professionals and members of the public are needed as their perceptions of the problems may diverge (Einfeld and Blomkamp 2021; Lewis et al. 2020). Hence, diversifying the sources of knowledge through the experiment can help policy-makers better understand and predict people's needs, perceptions and behaviors in real-life contexts (Clarke and Craft 2019; Einfeld and Blomkamp 2021; Lewis et al. 2020). In addition to refining policy solutions, stakeholders or end-users may also be involved in defining the policy *goals* and generating ideas for solutions (Clarke and Craft 2019; Einfeld and Blomkamp 2021; Lewis et al. 2020).

Design experiments share similarities with explorative pilots that are controlled only to a limited extent (Ansell and Bartenberger 2016). What makes the design experiments different is their explicitly *iterative* character. Thus, design experiments may provide dynamic and timely ways to change course during the experimentation process, therefore making them suitable for developing solutions for uncertain environments and complex problems. In tax policy, however, it is difficult to conceive of subjecting the stakeholders to varying tax rates, deductions, or exemptions in an iterative way. Given the importance of stable tax horizons in the investment and other decisions of individuals and businesses, such continuous changes and tweaks in the tax regime would be challenging.

Nevertheless, design experiments and the possibilities for iterative adjustments could be of value for developing novel tax policy solutions. For example, before introducing new taxes to quickly changing domains, the flexibility of the approach and insights from the stakeholders (Einfeld and Blomkamp 2021; Stoker and John 2009) may be crucial for the policymakers in assessing the feasibility of such taxes. As design experiments often seek to create prototypes and collect feedback on the potential responses to these policy measures (Clarke and Craft 2019; Einfeld and Blomkamp 2021), they could provide valuable insights about how the potential taxpayers, if they were subjected to such models, would react and adjust their behavior (Stoker and John 2009).

While RCTs require policy-makers to identify clear expectations about measurable impacts in order to commence the experiment, design experiments tend to have a "fuzzy front end", which allows for the "exploration of open-ended questions" (Clarke and Craft 2019, 9). That can be valuable in utilizing stakeholders' knowledge for making use of the tax system in solving new societal problems. By offering opportunities to test new approaches in an iterative way, design experiments offer a safe space for trying out novel solutions, reducing the fear of failure and hence promote innovation in policy design (Clarke and Craft 2019; Stoker and John 2009). Furthermore, by emphasizing the lived experiences of those affected by a policy design, such experiments can cater to the need to adapt the policy to different target populations as opposed to settling on a one-size-fits all approach (Clarke and Craft 2019, 7). Design experiments could be utilized in testing how potential tax payers would perceive information about new tax incentives before these changes are rolled out. The way tax incentives are presented can play a crucial role in how they are perceived by the taxpayers, and this, in turn can affect their potential tax behavior in the future (Werner and Riedl 2019). Complex tax regimes in particular can lead to weaker behavioral adjustments than expected by policy makers (e.g. Abeler and Jäger 2015). Design experiments can hence serve to assess the perceived complexity of a tax regime and help to mitigate potential distortions in the perception of the tax policy change.

In sum, design experiments could be valuable in understanding the reactions and opinions of policy target groups or developing new technological solutions that presume high inclusion of stakeholders in order to reach desired outcomes of tax policy. Nevertheless, design experiments also do face some crucial challenges in tax policy experimentation.

First, as continuous feedback and quick path adjustment is crucial during design experiments, they are not suitable in situations where effects of the action take a long time to appear or when there is a lack of control in different phases of experiment.

Second, design experiments tend to be applicable to small-scale policy issues rather than large-scale ones: they tend to be employed for "discrete service redesign projects" and in "exploratory work of scoping problems", rather than for the development of broader policy proposals or systemic reforms (Lewis et al. 2020, 113-114). It would be difficult to use them for assessing macroeconomic outcomes as the scope and intensity of design experiments (entailing immersion of experimenters and stakeholders) necessarily means a limited number of data points and the need to stay at the micro level.

Third, while the notions of "user centrism" inherent in design experiments may work well in the private sector context (where design thinking originates from), they may clash with the notions of rights and obligations of citizens, which prevail in tax policy. The stakeholders whose feedback is collected during the experiment may view taxes from their own narrow material point of view (with the goal to minimize their own tax burden), which may clash with broader policy goals. As Clarke and Craft (2019, 14) emphasize, user centrism may not be an easily applicable principle for contentious policy design choices, which include a broad range of users with conflicting needs and expectations, and thus require trade-offs between different values. The act of weighing these needs in tax policy is inherently political, subjective, normative, and ultimately falls to the accountable elected officials.

Conclusion

In light of the growing importance of experimental approaches in public policy, we proposed a binary question in the title of this article: to experiment or not to experiment in the field of tax policy? As the discussion above shows, the answer to that question is much more nuanced than a simple "yes" or "no". Although Abramowicz (2019) in his thought-provoking study advocated an extensive use of randomized experimentation in tax policy, our claims are considerably more cautious. Using experiments in tax policy does have a range of promises but also a wide spectrum of pitfalls.

On the one hand, experimental approaches have the potential to increase both the substantive quality of tax policy and the policy-making process. Experimental tax policy can potentially help policy makers alleviate some of the uncertainties and information asymmetries with regard to the actual impacts of new tax policy measures, avoid adverse effects, build consensus, and foster attention to consequences rather than ideological hunches in tax policy debates. On the other hand, however, tax policy experiments may give rise to ethical, legal and political challenges. Most of all, experimenting may be seen to violate the principles of horizontal equity, equal treatment and legal certainty. Tax policy is a field characterized by contentious policy choices, which engage a broad range of stakeholders with conflicting perceptions and needs. The act of balancing these perceptions is inherently complex, political and demands trade-offs between different value considerations. Altogether, the application of experimental approaches in the field of tax policy can be characterized by two core challenges that experimenters need to acknowledge and address. We label them a political challenge and a methodological challenge.

First, experimenting in tax policy may clash starkly – perhaps even more so than in many other policy fields – with the political nature of democratic decision-making. Most importantly, tax policy questions are likely to be profoundly influenced by the ideological leanings of policy makers and even define the core identity of some parties, which makes it challenging to test tax policy questions in a genuinely open way. Also, given that experimenting in tax policy would often entail materially benefitting or burdening (in a very clearly measurable way) some groups of taxpayers at the expense of others, such experiments may be more vulnerable than experiments in many other domains to legal challenges. Despite the main promise of experimental approaches to provide new knowledge in complex environments where various actors, institutions and technologies interact, the low tolerance of the tax policy field towards uncertainty and unpredictability may render the experimental approaches politically, ethically or legally unfeasible.

Second, methodologically, different experimental designs entail different strengths and promises, and there is no universal recipe to follow. The experimenters need to decide whether to prioritize causal explanations facilitated by RCTs or exploratory and collaborative aims fostered by other designs. While RCTs enable stronger causal conclusions, non-randomized pilots allow the testing of more holistic policy solutions, and design experiments offer more open-ended approaches and a stronger focus on stakeholder experience. Randomization inherent in RCTs might be a particularly challenging "political sell" in the field of tax policy, since it concerns the core rights and obligations of citizens. Both RCTs and policy pilots could face the accusations of violating horizontal equity – a key principle in tax policy. The user-centrism inherent in design experiments may clash with the notions of taxes as obligations.

Although we discussed the various designs as distinct options, the lines between them can in reality be somewhat blurred, and they could be used in a sequential manner (Bravo-Biosca 2020; Pomeranz and Vila-Belda 2019). For example, a design experiment could first help take a fresh look at some tax policy goals, followed by a simple non-randomized policy pilot that tests variants of a solution offered in an exploratory way, and an RCT could then zoom in on specific causal questions. Our recommendation is that policy-makers should remain open to a diversity of possible designs of experiments in tax policy. In particular, they should be aware of the dangers of conceiving of experimentation very narrowly, only in terms of RCTs – which seems to be the default given the tendency of the evidence-based policy-making movement to regard RCTs as the "gold standard" (Adkins and Ylöstalo 2018; Barrett and Carter 2010; Bedecarrats et al. 2019; Dalziel 2018; de Leao and Eyal 2019; Pearce and Raman 2014; Strassheim 2020).

In sum, there are considerable constraints for the scope of tax policy experimentation. Thus, while in some disciplines and perhaps even policy-fields experimentation could be the prominent *Zeitgeist,* in tax policy it is likely to remain on the margins. We conclude that in the field of tax policy, experimentation could be feasible in carefully crafted revenue neutral experiments, in phase-wise introductions of larger programs where limited resources can offer a justifiable reasoning for benefitting some taxpayers before others, and fostering stakeholder discussions in very new tax policy domains.

Acknowledgments

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

References

Abeler, J. and S. Jäger. 2015. "Complex Tax Incentives." *American Economic Journal*: *Economic Policy* 7(3), 1-28.

Abramowicz, M. 2008. "Randomized Legal Experimentation." *GWU Legal Studies Research Paper* 447. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1273875.

Abramowicz, M. 2019. "Tax Experimentation." *Florida Law Review* 71, 65-122.

Abramowicz, M., I. Ayres and Y. Listokin. 2010. "Randomizing Law." *University of Pennsylvania Law Review* 159(4), 929-1005.

Adkins, L. and H. Ylöstalo. 2018. "Experimental Policy, Price and the Provocative State." *Distinktion*: *Journal of Social Theory* 19(2), 152-169.

Akdogan, M.U. 2021. "Digital Service Taxes as the Fiscal Result of Digital Transformation." In U. Hacioglu and T. Aksoy (eds). *Financial Ecosystem and Strategy in the Digital Era*. Springer, 355-378.

Alm, J. and A. Malézieux. 2021. "40 Years of Tax Evasion Games: A Meta-Analysis." *Experimental Economics* 24(3), 699-750.

Anderson, A.B. 1975. "Policy Experiments: Selected Analytic Issues." *Sociological Methods & Research* 4(1), 13-30.

Ansell, C.K. and M. Bartenberger. 2016. "Varieties of Experimentalism." *Ecological Economics* 130, 64-73.

Bailey, S., K. Checkland, D. Hodgson, A. McBride, R. Elvey, S. Parkin, K. Rothwell and D. Pierides. 2017. "The Policy Work of Piloting: Mobilising and Managing Conflict and Ambiguity in the English NHS." *Social Science & Medicine* 179, 210-217.

Ballard-Rosa, C., L. Martin and K. Scheve. 2017. "The Structure of American Income Tax Policy Preferences." *The Journal of Politics* 79(1), 1-16.

Barrett, C.B. and M.R. Carter. 2010. "The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections." *Applied Economic Perspectives and Policy* 32(4), 515-548.

Bauknecht, D., T.S. Bischoff, K. Bizer, M. Führ, P. Gailhofer, D.A. Heyen, T. Proeger, and K. von der Leyen. 2020. "Exploring the Pathways: Regulatory Experiments for Sustainable Development – An Interdisciplinary Approach." *Journal of Governance and Regulation* 9(3), 49-71.

Bedard, P.O. and M. Ouimet. 2012. "Cognizance and Consultation of Randomized Controlled Trials among Ministerial Policy Analysts." *Review of Policy Research* 29(5), 625-644.

Bedecarrats, F., I. Guerin and F. Roubaud. 2019. "All that Glitters is not Gold: The Political Economy of Randomized Evaluations in Development." *Development and Change* 50(3), 735-762.

Bell, S.H. and L.R. Peck. 2016. "On the Feasibility of Extending Social Experiments to Wider Applications." *Journal of MultiDisciplinary Evaluation* 12(27), 93-112.

Bravo-Biosca, A. 2020. "Experimental Innovation Policy." *Innovation Policy and the Economy* 20(1), 191-232.

Burtless, G. 1995. "The Case for Randomized Field Trials in Economic and Policy Research." *Journal of Economic Perspectives* 9(2), 63-84.

Carmines, E.G., M.J. Ensley and M.W. Wagner. 2012. "Who Fits the Left-Right Divide? Partisan Polarization in the American Electorate." *American Behavioral Scientist* 56(12), 1631-1653.

Checkland, K., J. Hammond, A. Coleman, J. Macinnes, R. Mikelyte, S. Croke, J. Billings, S. Bailey, and P. Allen. 2021. "'Success' in Policy Piloting: Process, Programs, and Politics." *Public Administration*. Published ahead-of-print 1-18. https://doi.org/10.1111/padm.12790.

Clarke, A. and J. Craft. 2019. "The Twin Faces of Public Sector Design." *Governance* 32(1), 5-21.

Cook, T.D. 2002. "Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community Has Offered for Not Doing Them." *Educational Evaluation and Policy Analysis* 24(3), 175-199.

Cotterill, S. and L. Richardson. 2010. "Expanding the Use of Experiments on Civic Behavior: Experiments with Local Government as a Research Partner." *The Annals of the American Academy of Political and Social Science* 628(1), 148-164.

Dalziel, M. 2018. "Why are there (Almost) No Randomised Controlled Trial-Based evaluations of Business Support Programmes?" *Palgrave Communications* 4(1), 1-9.

Deaton, A. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2), 424-55.

Druckman, J.N., D.P. Green, J.H. Kuklinski and A. Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4), 627-635.

Einfeld, C. and E. Blomkamp. 2021. "Nudge and Co-Design: Complementary or Contradictory Approaches to Policy Innovation?" Published ahead-of-print in *Policy Studies*. https://doi.org/10.1080/01442872.2021.1879036.

Elkins, D. 2006. "Horizontal Equity as a Principle of Tax Theory." *Yale Law & Policy Review* 24, 43-90.

Ettelt, S. and Mays, N. 2015. "RCTs: How Compatible Are They with Policy-Making?" *British Journal of Healthcare Management* 21(8), 379-382.

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.

Farrelly, R. 2008. "Policy on Trial: Randomised Trials Are the Best Tool We Have for Finding Out if Policies Really Work." *Policy: A Journal of Public Policy and Ideas* 24(3), 7-12.

Hallsworth, M. 2014. "The Use of Field Experiments to Increase Tax Compliance." *Oxford Review of Economic Policy* 30(4), 658-679.

Haynes, L., B. Goldacre and D. Torgerson. 2012. "Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials." Cabinet Office-Behavioural Insights Team. Available

at http://dx.doi.org/10.2139/ssrn.2131581.

Heilmann, S. 2008. "Policy Experimentation in China's Economic Rise." *Studies in Comparative International Development* 43(1), 1-26.

Heldeweg, A.M. 2015. "Experimental Legislation Concerning Technological & Governance Innovation: An Analytical Approach." *The Theory and Practice of Legislation* 3(2), 169-193.

Hughes, S., S. Yordi, L. and Besco. 2020. "The Role of Pilot Projects in Urban Climate Change Policy Innovation." *Policy Studies Journal* 48(2), 271-297.

Huitema, D., A. Jordan, S. Munaretto and M. Hilden. 2018. "Policy Experiments: Core Concepts, Political Dynamics and Impacts." *Policy Sciences* 51(2), 143-159.

John, P. 2014. "Policy Entrepreneurship in UK Central Government: The Behavioural Insights Team and the Use of Randomized Controlled Trials." *Public Policy and Administration* 29(3), 257-267.

Jones, R. and M. Whitehead. 2018. "'Politics Done like Science': Critical Perspectives on Psychological Governance and the Experimental State." *Environment and Planning D: Society and Space* 36(2), 313-330.

Karvonen, A. and B. Van Heur. 2014. "Urban Laboratories: Experiments in Reworking Cities." *International Journal of Urban and Regional Research* 38(2), 379-392.

Kay, J.A. 1990. "Tax Policy: A Survey." *The Economic Journal* 100(399), 18-75.

Klitgaard, M.B., G. Schumacher and M. Soentken. 2015. "The Partisan Politics of Institutional Welfare State Reform." *Journal of European Public Policy* 22(7), 948-966.

Ko, K. and K. Shin. 2017. "How Asian Countries Understand Policy Experiment as Policy Pilots?" *Asian Journal of Political Science* 25(3), 253-265.

Kvangraven, I.H. 2020. "Nobel Rebels in Disguise: Assessing the Rise and Rule of the Randomistas." *Review of Political Economy* 32(3), 305-341.

de Leao, L.S. and G. Eyal. 2019. "The Rise of Randomized Controlled Trials (RCTs) in International Development in Historical Perspective." *Theory and Society* 48(3), 383-418.

Lee, C. and L. Ma. 2020. "The Role of Policy Labs in Policy Experiment and Knowledge Transfer: A Comparison across the UK, Denmark, and Singapore." *Journal of Comparative Policy Analysis: Research and Practice* 22(4), 281-297.

Lee, S.-Y., Y.-R. Jung and E. Lee. 2009. "An Exploratory Study on Pilot Projects: The Ministry for Health, Welfare, and Family Affairs in Korea." *The Korean Journal of Public Administration* 47(4), 279-305.

Levitt, S.D. and J.A. List. 2011. "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments." *American Economic Journal: Applied Economics* 3(1), 224-238.

Lewis, J.M., 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.

McFadgen, B.K. 2019. "Connecting Policy Change, Experimentation, and Entrepreneurs." *Ecology and Society* 24(1), 30.

McFadgen, B. and Huitema, D. 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(2), 161-187.

McGann, M., E. Blomkamp and J.M. Lewis. 2018. "The Rise of Public Sector Innovation Labs: Experiments in Design Thinking for Policy." *Policy Sciences* 51(3), 249-267.

Mascagni, G. 2018. "From the Lab to the Field: A Review of Tax Experiments." *Journal of Economic Surveys* 32(2), 273-301.

Millo, Y. and J. Lezaun. 2006. "Regulatory Experiments: Genetically Modified Crops and Financial Derivatives on Trial." *Science and Public Policy* 33(3), 179-190.

Moffitt, R.A. 2003. "The Negative Income Tax and the Evolution of U.S. Welfare Policy." *Journal of Economic Perspectives* 17(3), 119-140.

Musgrave, R.A. 1990. "Horizontal Equity, Once More." *National Tax Journal* 43(2), 113-122.

Nair, S. and M. Howlett. 2016. "Meaning and Power in the Design and Development of Policy Experiments." *Futures* 76, 67-74.

Oakley, A. 1998. "Public Policy Experimentation: Lessons from America." *Policy Studies* 19(2), 93-114.

Osterloh, S. and M. Debus. 2012. "Partisan Politics in Corporate Taxation." *European Journal of Political Economy* 28(2), 192-207.

Pearce, W. and S. Raman. 2014. "The New Randomised Controlled Trials (RCT) Movement in Public Policy: Challenges of Epistemic Governance." *Policy Sciences* 47(4), 387-402.

Peters, B.G. 1998. "The Experimenting Society and Policy Design." In W. Dunn (ed.). *The Experimenting Society. Essays in Honor of Donald T. Campbell*. New Brunswick, NJ: Transaction, 125-140.

Philipsen, S., E.F. Stamhuis and M. de Jong. 2021. "Legal Enclaves as a Test Environment for Innovative Products: Toward Legally Resilient Experimentation Policies." *Regulation & Governance* 15(4), 1128-1143.

Pomeranz, D. and J. Vila-Belda. 2019. "Taking State-Capacity Research to the Field: Insights from Collaborations with Tax Authorities." *Annual Review of Economics* 11, 755-781.

Ranchordas, S. 2013. "The Whys and Woes of Experimental Legislation." *The Theory and Practice of Legislation* 1(3), 415-440.

Ranchordas, S. 2015a. "Innovation Experimentalism in the Age of the Sharing Economy." *Lewis & Clark Law Review* 19, 871-924.

Ranchordas, S. 2015b. "Time, Timing, and Experimental Legislation." *The Theory and Practice of Legislation* 3(2), 135-139.

Ranchordas, S. 2015c. "Innovation-Friendly Regulation: The Sunset of Regulation, the Sunrise of Innovation." *Jurimetrics* 5(2), 201-224.

Rangoni, B. and J. Zeitlin. 2021. "Is Experimentalist Governance Self-Limiting or Self-Reinforcing? Strategic Uncertainty and Recursive Rulemaking in European Union Electricity Regulation." *Regulation & Governance* 15(3), 822-839.

Shome, P. 2021. *Taxation History, Theory, Law and Administration*. Springer.

Stein, H., S. Cunningham and P. Carmody. 2021. "The Rise of 'Behavioral Man': Randomized Controlled Trials and the 'New' Development Agenda." *Human Geography* 14(1), 62-75.

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 P. John. 2009. "Design Experiments: Engaging Policy Makers in the Search for Evidence about What Works." *Political Studies* 57(2), 356-373.

Strassheim, H. 2020. "De-Biasing Democracy. Behavioural Public Policy and the Post-Democratic Turn." *Democratization* 27(3), 461-476.

Tõnurist, P., R. Kattel and V. Lember. 2017. "Innovation Labs in the Public Sector: What They Are and What They Do?" *Public Management Review* 19(10), 1455-1479.

Van Gestel, R. and G. Van Dijck. 2011. "Better Regulation through Experimental Legislation." *European Public Law* 17(3), 539-533.

Van der Heijden, J. 2018. "From Leaders to Majority: A Frontrunner Paradox in Built-Environment Climate Governance Experimentation." *Journal of Environmental Planning and Management* 61(8), 1383-1401.

Van der Heijden, J. and S.H. Hong. 2021. "Urban Climate Governance Experimentation in Seoul: Science, Politics, or a Little of Both?" *Urban Affairs Review* 57(4), 1115-1148.

Voß, J.P. and A. Simons. 2018. "A Novel Understanding of Experimentation in Governance: Co-producing Innovations between 'Lab' and 'Field'". *Policy Sciences* 51(2), 213-229.

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. Ulu (eds.). *Handbook of Research Methods and Applications in Experimental Economics.* Cheltenham: Edward Elgar Publishing, 295-311.

Zurbriggen, C. and M.G. Lago. 2019. "An Experimental Evaluation Tool for the Public Innovation Lab of the Uruguayan Government." *Evidence & Policy* 15(3), 437-451.

Ringa Raudla (ringa.raudla@taltech.ee) is Professor of Fiscal Governance, Ragnar Nurkse Department of Innovation and Governance, Tallinn University of Technology, Estonia. Her main research interests are fiscal policy, public budgeting, public administration reform, and institutional economics. Her recent publications include articles in *Public Administration Review*, *Governance*, *Public Administration*, *Policy Studies Journal*, *the American Review of Public Administration*, *Journal of Public Policy*, *Public Money & Management*, *Public Budgeting & Finance, Public Management Review, International Review of Administrative Sciences*, and *Journal of Comparative Policy Analysis*. She has also worked as a consultant for various governmental organizations, including the Ministry of Finance and the National Audit Office in Estonia, and international organizations, including the World Bank. Ringa Raudla received the Estonian national research award in social sciences in 2018. She is an editorial board member of seven different journals in the fields of public administration, public management, governance and economics.

Külli Sarapuu (kylli.sarapuu@taltech.ee) is Associate Professor of Public Sector Management and Organization, Ragnar Nurkse Department of Innovation and Governance, Tallinn University of Technology, Estonia. Before joining TalTech, she worked for the Estonian Government Office and the Ministry of Research and Education. Külli's research focuses on public sector organization, civil service systems, public sector coordination, and the governance of small states. She has published numerous book chapters and articles on these topics, including in the journals *Public Management Review*, *International Review of Administrative Sciences* and *International Journal of Public Administration*.

Egert Juuse is a Research Fellow, Ragnar Nurkse Department of Innovation and Governance, Tallinn University of Technology, Estonia. He is also the head of two Master level programs: Technology Governance and Digital Transformation (MA) and E-Governance and Public Sector Innovation (MSc). He has received a BA, MA and PhD in Technology Governance from Tallinn University of Technology. His research interests include various aspects related to economic development with a particular focus on global value chains; financial policies (incl. Europeanization); financialization and financial crises; innovation and industrial policies. He has also worked as an advisor to a Minister at the Ministry of Economic Affairs and Communications, Estonia.

Nastassia Harbuzova is a Doctoral Student and a Junior Research Fellow at the Ragnar Nurkse Department of Innovation and Governance, Tallinn University of Technology. Her research concentrates on macroeconomic policymaking from both ideational and implementation perspectives. Currently, she is focused on studying how policy learning occurs in field experiments undertaken in different institutional contexts. Her educational background includes Economic Policy, Mathematical Methods in Economics and Intellectual Property Management. Previously she worked as a Lecturer at the Belarusian State University as well as co-led a student science club, CLIO BSU, focusing on research in hetero- and orthodox economics. In 2021 she published a taskbook, *Macroeconomics*, for undergraduates utilized by BSU.

Kerli Onno is an industrial PhD student at the Nurkse Department. Her research focuses on the value of data capabilities in public sector organizations. As she works as the Head of the Analysis Unit at the Estonian Tax and Customs Board, she is deeply interested in the ways data capabilities support public sector organizations in adapting to the changing environments and in developing dynamic ways of operating to provide long-term public value. In addition, she is exploring various experimental approaches in public sector organizations and is involved in a research project, *Experimental approaches and institutional innovations in the domains of fiscal and financial policy*. Kerli Onno holds an MA in Communication Management (*cum laude*) from Tallinn University and an MA in Internal Security from the Estonian Academy of Security Sciences with a focus on media studies and strategic communication in national defence.

Johanna Vallistu is a PhD student at the Ragnar Nurkse Department of Innovation and Governance, Tallinn University of Technology. She also works as an expert at the Foresight Centre at the Parliament of Estonia. At the Foresight Centre, Johanna has led the foresight projects of Labor Market 2035, Financial Well-being of the Future Elderly and currently leads the project on the Future of Data Freedom. Johanna's main interests lie in the future social insurance system and how digitalization impacts governance.

Aleksandrs Cepilovs is a research fellow (part-time) at the Ragnar Nurkse Department of Innovation and Governance. He is also a research officer at the European Union Agency for the Operational Management of Large-Scale IT Systems in the Areas of Freedom, Justice and Security (eu-LISA). His research interests include innovation policy, policy transfer and policy learning, public sector innovation and e-Governance. Most recently, Aleksandrs has been involved in a research project focusing on experimental approaches and institutional innovations in the domains of fiscal and financial policy.