Leveraging Field Experiments to Shape Environmental Policy

Af
Christina Gravert*
University of Copenhagen

Keywords: policy experiments, environmental policy
JEL codes: C9, Q5

Abstract

Deciding which environmental policies to implement is crucial for governments and essential for protecting our planet. Numerous policy questions from energy use, to meat consumption, to public transport usage require empirical answers. Ideally, policymakers would rely on causal evidence to design policies. However, in environmental policy this is rarely done. This paper advocates for policy experiments as a feasible and valuable approach to designing effective environmental policies ex-ante. I address common objections to policy experiments, such as feasibility, lack of skills and possible objections of voters and show how they can be overcome. I use examples from my research to illustrate the potential of policy experiments in answering critical environmental questions and promoting evidence-based policy making.

* Department of Economics and Center for Economic Behavior and Inequality (CEBI), University of Copenhagen. Email: cag@econ.ku.dk. The activities of CEBI are financed by the Danish National Research Foundation, Grant DNRF134.

1. Introduction

Deciding which environmental policies to implement is a crucial task for governments and a necessary step in protecting our planet. Numerous environmental policy questions need answers: How do time-of-use tariffs affect consumer electricity consumption? To what extent will a tax on meat reduce red meat consumption? How significantly will subsidies for public transport increase its usage? How do energy labels for houses influence renovation efforts? What types of communication strategies enhance public acceptance of green policies?

In an ideal world, policymakers would find answers to these questions and design policies based on evidence of what works. Unfortunately, this is rarely the case. A significant portion of environmental policy set by ministries and governmental agencies results from political negotiations, engineering calculations, the implementation of international laws and targets, or even intuition, rather than scientific evidence.

In this paper, I argue that policy experiments are a valuable and feasible addition to the design of environmental policy. What do I mean by a policy experiment? To estimate the effect of a policy accurately, it is essential to implement the policy on a randomly chosen subset of units, such as individuals, regions, or firms. This approach allows for the observation of counterfactual outcomes by comparing the control group (those not subjected to the policy) with the treated group (those subjected to the policy). By analyzing the differences in outcomes between these two groups, we can determine the policy's impact. If certain conditions are met, this method provides a precise and unbiased estimate of the policy's effect, thereby enabling policymakers to make informed decisions about whether to implement the policy on a broader scale.

A policy experiment is conducted ex-ante, that is, before the policy is implemented. Consider the example of a missed opportunity for a policy experiment and how it could have been improved. In 2023, Denmark introduced time-of-use tariffs intending to reduce household electricity use during peak times. Grid tariffs during 17-21 in the evening, when electricity demand is high, are almost ten times higher than tariffs during 0-6 at night when demand is low. From January 1, 2023, all local distribution companies had to implement these tariffs. Did these tariffs have the intended effect on consumption? Are they set at the ideal rates? We don't know because we don't observe the counterfactual outcome.

What would have happened under a different tariff structure? A policy experiment could have provided us with this answer. Instead of rolling out the new tariff to everyone, some distribution areas could have been randomly selected to receive the new tariff, some to stay on the old tariff, and some to receive a third version of the tariffs. After a few weeks, Energinet, the Danish grid operator, would then have been able to analyze whether electricity consumption varied between the three groups. Based on this evidence, policymakers could have deci-

ded on the most effective tariff to reduce peak consumption or, if the tariffs had no effect, chosen another policy tool to shift consumption. The administrative costs of such an ex-ante policy experiment would be minimal compared to the costs of maintaining an ineffective policy. Without this knowledge, it is also unclear whether the tariffs should be reformed and how.

The idea of using experiments to inform policy is not novel. In the US, several social experiments on employment programs, electricity pricing, job training programs, and housing allowances were conducted as early as the 1960s. In 2019, Esther Duflo, Abhijit Banerjee, and Michael Kremer received the Nobel Prize in Economics for their work on field experiments in low- and middle-income countries. Even in Denmark, the Danish Agency for Labour Market and Recruitment (STAR) has been conducting labor market activation experiments since 2005, aiming to improve the effectiveness of employment services. As *Kreiner & Svarer (2022)* write: »There has been a strong focus in recent decades on evidence-based policy-making in Denmark's active labor market policy. The goal is that decisions on how to design the policy and on the amount of resources to use rely as far as possible on cost-benefit analyses based on high-quality empirical evidence. [...] The systematic use of randomized control trials to evaluate the impact of the active labor market policies is a rather unique feature of the Danish labor market policy. The randomized control trials have the additional advantage that they provide a natural set ting for evaluating the cost-effectiveness of the programs.« One example of these policy experiments is the »Quickly Back to Work« series, which began in 2005. This series of experiments aimed to improve the effectiveness of active labor market policies by providing job seekers with targeted support and resources. The Danish Economic Council has calculated that the program has saved around 15,000 DKK ($2500) per unemployed person in the experiment (*Kreiner & Svarer 2022*). Similarly, since 2013, the TrygFonden's Center for Child Research at Aarhus University has been working closely with municipalities and ministries to test policies related to child development. One example of their work is an intervention tested in 122 Danish pre-schools that aimed to increase school readiness for pre-school children from disadvantaged backgrounds (*Jensen & Sjö 2024*). 73 pre-schools were allocated to the intervention group, and 49 were allocated to the control group. Over 5000 children participated in the experiment. Interestingly, the study found very little effect of the quite extensive program on the intended outcomes. Without the experiment, significant public resources might have been spent on a program that did not achieve the desired effects

These examples demonstrate that conducting policy experiments to inform policy is neither novel nor impossible and that significant government resources can be allocated more optimally if policymakers have evidence on the effectiveness of programs and policies. In the following sections, I will explain what a policy experiment is and discuss common arguments against running policy experiments. Using examples from my own research, I will provide counterarguments and

show possible solutions to how important environmental policy questions can be answered using an experimental approach

2. What is a policy experiment?

A policy experiment is a systematic approach used by governments and institutions to test and evaluate the effectiveness of policy interventions before they are implemented on a larger scale. To understand why a policy experiment is so effective in providing causal evidence, I will briefly describe the potential outcomes framework slightly adapted from *List et al. (2011*).

We consider the outcome Y_i of subject i under treatment (new policy) and control (old policy), $T = 1$ and $T = 0$, respectively. We assume that it can be modeled as a function of observable variables X_i, an unobserved person-specific effect α_i, an average treatment effect $\bar{\tau}$, and a noise term ε_i, which is assumed to be independent and identically distributed (i.i.d.)

$$Y_{iT} = \alpha_i + \beta X_i + \bar{\tau} T + \varepsilon_i$$

The average treatment effect can then be defined as

$$\bar{\tau} = E(Y_{i1} - Y_{i0}) = E(Y_{i1}) - E(Y_{i0})$$

The identification problem is that we can only observe $E(Y_{i1}|\, T = 1)$ and $E(Y_{i0}|\, T = 0)$, where $T = 1$ and $T = 0$ for a given i. Because it is impossible to observe unit i in both states, i.e., we cannot observe $E(Y_{i1}|\, T = 0)$and $E(Y_{i0}|\, T = 1)$, it is necessary to construct a proper counterfactual. What would have happened if individual i would/ would not have been treated? If we can assume that the individuals in both control and treatment group are identical on average, then we have a proper counterfactual. However, if the propensity to receive treatment is correlated with any of the unobserved variables, for example, because some subjects have self-selected into a policy, then the estimate of the average treatment effect is biased since

$$\hat{\tau} = E(Y_{i1}|\, T = 1) - E(Y_{i0}|\, T = 0) \neq E(Y_{i1}) - E(Y_{i0}) = \bar{\tau}$$

The only way, we can be sure that that the groups are equivalent in all respects except for the treatment (also on unobservable traits) is if we randomize them into groups. If we randomly assign individuals to treatment, for example, by conducting a lottery, we can generally be confident that the treatment is not correlated with any observed or unobserved variables. When we compare those who received the treatment with those who did not, we can estimate the unbiased average treatment effect. This can usually be done using a simple ordinary least

squares regression or even simple tests of means, making it econometrically far easier than running more complex econometric models that require numerous assumptions and a large set of control variables.

Clearly, policy experiments run in the field will face challenges. Will the sample size be large enough to assume that, on average, the groups are the same on all characteristics? Will everyone comply with the treatment? Will everyone see the treatment? Can we randomize at the level of the unit of observation? Will the treatment and control groups influence each other in some way? While these are all valid questions that need to be addressed when running experiments, policymakers can rest assured that solutions to these problems exist. This article is not a manual on how to conduct experiments (I teach that course at KU every spring and have already published a guideline for policy makers to evaluate nudges (*Gravert & Carlsson 2019*). Instead, I will focus on the barriers that prevent policy experiments from being attempted at all. These are the more challenging problems to handle.

In this paper, I use the terms policy experiment, field experiment, and randomized controlled trial (RCT) interchangeably to refer to an experiment carried out in a natural setting with real-world incentives. Often, this means that participants are unaware they are part of an experiment, as is the case with natural field experiments (*Harrison & List 2004*). In such cases, the assumptions presented above can be taken as given. Sometimes, it is unavoidable to inform participants that they are part of an experiment for practical or ethical reasons. These experiments are then called framed field experiments. They still occur in a real-world setting with real stakes, but participants know they are being observed and might be asked to fill out a survey. In these cases, it is important to be aware of who chooses to participate in the experiment to understand the external validity of the research findings.

3. Why are field experiments so rarely done to inform environmental policy?

So why is it that we see so few policy experiments in environmental policy? In the following sub-sections I will go through a number of arguments that might explain why policy experiments are so rare. I focus on Denmark, but according to a recent paper in Science written by more than a dozen economists from around the world, many of the challenges and solutions apply more generally (*Ferraro et al. 2023*). Further, most of the arguments will apply to policy experiments in all areas of public policy.

3.1. There are practical, financial and legal barriers

As with any policy domain, there may be legal or political barriers to changing policies or implementing a new policy for only a subset of the population and for

a limited time. Instead of experimenting with a new policy, I therefor recommend to experiment with features of an existing program over which policymakers already have control. The following experiment demonstrates that it is simple and cost-effective to improve policies that are already in place.

Skånetrafik, the public transport agency in the region of Skåne in Sweden, had been running the same public transport campaign for several years. Each month, they collected the addresses of people who had moved to or within Skåne. They then sent these households a letter, welcoming them to the neighborhood and offering a two-week free public transport card, along with information about the nearest bus or tram stop. They knew that citizens appreciated the campaign and that many people requested the free travel cards. However, they did not know whether their campaign increased the likelihood of people using public transport or whether a different campaign could be more effective.

Fortunately, they were open to experimentation. We conducted several experiments, the main one published in *Gravert & Collentine (2021)*. We selected over 14,000 individuals who had recently moved to or within Skåne and randomly assigned them to three groups. One group received the standard letter with the offer of a two-week free travel card. Another group received the same letter and offer, but with a social norm nudge stating truthfully that »the majority of their neighbors occasionally use public transport.« The third group received an offer of a four-week travel card. We found that the nudge had no effect on uptake or activation of the card, but the four-week card significantly increased uptake and long-term public transport use, even several months later, compared to the two-week card. Ideally, we would have measured the effect of no incentives (sending them a card without any free travel time), but since the experiment was designed to not make anyone worse off than if no experiment had taken place, everyone received at least the two-week card. This example also shows that the financial feasibility of an experiment is less problematic than often assumed. The administrative costs of conducting the experiment compared to running the campaign as usual were minor, mostly involving printing three different flyers instead of one. The additional marginal costs for providing a four-week versus a two-week travel card were also small, as this was only the lost revenue from those who would have otherwise paid for their tickets during the extra two weeks. Given the significant increase in purchases of monthly tickets by the four-week group compared to the two-week group, these costs were more than compensated for in the following months. The data we used was already being collected by Skånetrafik. We only needed to track which individual had been randomized into which group. An important outcome was that the nudge, which had proven effective in other settings (see discussion in section 3.5), had zero impact on transport behavior. Without the experiment, they might have implemented the nudge under the belief that it would be effective. This experiment is a good example of how existing

campaigns can be leveraged for simple experimentation to overcome financial, practical, and legal barriers.

There can, of course, be policies that use incentives which are more expensive to test. However, the financial viability of a policy should always be considered before testing. There is no point in testing an intervention that is so expensive that it cannot be scaled up if it proves effective. If the experiment is well-designed and utilizes established processes and data already being collected, the financial costs of an experiment are negligible compared to the benefits of knowing whether the policy has the intended effects.

Sometimes, we might be interested in testing policies for which there is no established process and that have not yet been implemented. Can policy experiments help in this case? For example, before implementing a carbon tax on consumer products, we might want to know the demand effects of such a tax and determine the appropriate tax level to set the right incentives.

Together with my co-authors, we explored this question.[1] To test how a carbon tax on everyday consumer products would affect consumption, we conducted an online randomized controlled trial with representative sample of 3,000 UK citizens. Participants were asked to make grocery shopping decisions under one of three price schemes to which they were randomly assigned: a carbon tax, an import tax, or a price increase of equal size to the tax. We implemented several policy scenarios to test different levels of the tax. To make the experiment as realistic as possible, we used the carbon tax revenue to decommission emissions certificates from the EU emissions trading scheme, and we used the import tax revenue to buy government bonds and return them to the UK government. A random subset of participants also received their shopping choices sent to their homes. This type of experiment can be classified as a framed field experiment. Participants are aware that they are in an experiment, but the setting and stakes are as realistic as possible. We found that while a tax generally decreases consumption of high-carbon goods, for climate-concerned individuals, the tax seems to crowd out their intrinsic motivation to consume low-carbon goods. When the tax is zero, climate-concerned individuals are more likely to choose low-carbon goods, but once a tax is levied, some switch to high-carbon goods. This behavior aligns with the concept of moral licensing. By paying the tax, they can alleviate the guilt of consuming high-carbon goods. Without a tax, climate concerns motivate them to choose low-carbon goods to avoid feeling guilty about their consumption. This behavior contrasts sharply with traditional price theory. Depending on the share of climate-concerned individuals in the population, crowding out might reduce the tax's efficiency, especially at lower tax levels. This example shows that relying on economic theory for policy making can be problematic when behavioral respon-

1. Working paper is not yet available for citation.

ses deviate from the rational actor model used in most environmental economic theory.

While our online experiment provides novel and interesting insights into the effect of carbon taxation, it would be useful to test the same hypothesis in a setting where participants are unaware they are part of an experiment to avoid self-selection and experimenter demand effects.[2] Ideally, a carbon tax would be added to several consumer products in supermarkets to measure the effect on demand compared to the same products in supermarkets without the added tax or compared to similar products. The German retailer Penny did something similar in 2023. They changed the prices of a handful of consumer products to their »real prices«, considering their climate and social impact (*Tagesschau 2024*). Unfortunately, the initiative was only implemented as a marketing campaign and not as a randomized controlled trial, making it difficult to draw conclusions from the initiative. Nevertheless, it showed that it would be possible to do something similar as a policy experiment to gather additional knowledge about the effects of a carbon tax on consumer behavior. In section 4, I present an experiment I conducted with supermarkets in Sweden, which shows how one could turn an initiative like this into a proper policy experiment.

Finally, policy makers might be concerned about the time an experiment might take before a policy can be implemented. While it might be ideal from an academic standpoint to have a long experimental period and follow-up, this is normally not necessary for policy making. The experiment with Skånetrafik lasted a few weeks and the data for the carbon tax experiment was collected in two days. Most experiments presented in this article only lasted a few weeks, at most. For many environmental problems the culprit is human behavior, so if we measure human behavior (driving a car), rather than the outcomes the behavior produces (pollution levels) the desired changes can be measured much quicker.

3.2. We don't know how

Ferraro et al. (2023) suggest that one reason for the lack of policy experiments in environmental policy is that the field is dominated by lawyers, engineers, and natural scientists. These professionals are not exposed to the idea of experimentation to the same extent as health and social scientists and, therefore, might not anticipate complex human responses to seemingly straightforward policies and incentives. Even among economists working in governmental organizations, there is often little to no knowledge of how to conduct randomized controlled trials for policy making. While RCTs are discussed as a method in several Master's level courses in economics programs, it was only in 2021 that I developed the first full-

2. Experimenter demand effects can be problematic if participants try to act in a way they think the experimenter wants. While these effects are often small, they can bias the results (De Quidt et al. 2018).

semester course on conducting field experiments at the University of Copenhagen. It is, therefore, not surprising that policymakers are not trained in conducting policy experiments and do not consider them part of their toolkit. Instead, there is a strong focus on difference-in-differences estimation or even simple correlational techniques. Importantly, as described in section 2, analyzing results from a well-conducted policy experiment is usually easier than the more advanced econometric methods necessary for correcting ex-post for a lack of proper random assignment. Thus, conducting and evaluating policy experiments is not technically more difficult than any other econometric exercise. It only requires more foresight.

The work of STAR and the TrygFonden's Center for Child Research demonstrates that experimental skills can be acquired by policymakers or, at least, that policy experiments can be conducted in collaboration with academics or consultants who have the necessary skills. In both cases, academics with expertise in conducting RCTs approached policymakers to propose running policy experiments to learn about what works.[3] There is nothing about job seekers or young children that makes them particularly well-suited for policy experiments, nor is there anything about employees at these agencies that makes them particularly well-equipped to run policy experiments. Importantly, even for academics, many skills are learned through practice, so fostering a culture of experimentation in policy domains builds skills for all participants and creates case studies for future reference. If there is an interest in learning how to conduct policy experiments and acquiring skills, there are several opportunities for partnerships with academics, learning from colleagues in other ministries, or collaborating with behavioral insights teams from other countries and fields. Many countries, such as the UK and the US, and international organizations like the UN, the World Bank, and the World Resources Institute, as well as large companies like the Commonwealth Bank, Google, and Walmart, have established formal behavioral insight teams to help other divisions implement insights from behavioral science and conduct randomized controlled trials. Institutions such as the Abdul Latif Jameel Poverty Action Lab (J-PAL) provide free resources and training to ensure that policy worldwide is informed by scientific evidence.

3.3. We don't have the necessary data

Another challenge with running field experiments for policy is the lack of relevant data available to measure outcomes. Collecting new data through surveys, observation, or by setting up new types of administrative data can be very expen-

3. An example of this is Michael Rosholm, who initiated much of the early labor market activation research in Denmark and founded the TrygFonden's Center for Child Research. The importance of one person in changing how something is done will be discussed more in section 3.5.

sive and is usually not a feasible option for policymakers. Often, data on a program can only be collected from those who participate (e.g., an energy audit), which creates significant challenges when there is low uptake and does not allow for comparison with those who have not opted into the program.

However, in Nordic countries such as Denmark, significant investments in data infrastructure have already been made. Using ten-digit identification numbers, CPR (for individuals) or CVR (for firms), data on individual units can be combined from hundreds of different data sources covering all areas of life, from cradle to grave. This data structure is unparalleled globally and provides the Nordic countries with an opportunity to conduct policy research at a level that cannot be matched elsewhere (*Frank 2000*). While thousands of studies have been conducted in the fields of health, labor, and education, the data has yet to be fully utilized for answering environmental policy questions.

Denmark was the first country in the world to achieve full coverage of electricity smart meters nationwide. Now, Energinet, the Danish grid operator, centrally collects hourly electricity production and consumption data for every household. Through Statistics Denmark, this meter data can be matched with dwellings and personal characteristics of the inhabitants. Most field experiments on electricity usage in other countries, such as the US, Germany, or the UK, have been conducted in collaboration with individual utilities and their selected set of customers. Such an approach reduces the external validity of the study and often results in smaller than desirable sample sizes. My recent work on consumer switching in electricity markets could not have been done without the Danish data infrastructure (*Gravert 2024*). I examined switching between electricity suppliers, which requires observing a random subset of the full population and all electricity suppliers.

In May 2022, I invited a random 100,000 individuals from the Danish population to participate in a survey on electricity markets. They received the invitation through e-boks, the Danish digital postbox (e-boks). Around 20 % of the invited individuals responded. The survey asked several questions about the respondents' electricity contracts, preferences, and knowledge about the electricity market. At the end of the survey, participants were randomized into three groups. The information group received information about how to switch electricity suppliers and how much they could expect to save. The broker group received the same information and the opportunity to use a broker to switch them to the cheapest supplier. The control group received a link to a website where they could compare suppliers. I asked the respondents whether they planned to switch suppliers in the next three months and recorded whether they clicked on the link to elpris.dk. This is where most other studies need to stop, as they cannot connect survey data with choice data. However, with the Danish data infrastructure, using individual, anonymized CPR numbers, I was able to connect the survey data with both background information from the national registries, such as

income, education level, and household size, as well as with the new smart meter data supplied by Energinet. This allowed me to conduct a heterogeneity analysis on how the information treatments affected choices, but more importantly, to measure whether the treatments affected actual switching. This paper is a hybrid between a survey and a field experiment, as there was self-selection into the survey. This meant that those least interested in electricity markets likely did not participate. On the positive side, the survey provided many insights, such as personality traits and perceptions of the market, that I could not have extracted from the administrative data. My survey experiment is, to the best of my knowledge, the first randomized controlled survey experiment conducted on a random subset of the population that also measures real-world outcomes.

To test the effect of information interventions on switching in a natural field experiment without conducting a survey, policymakers could use the e-boks infrastructure to send randomly assigned different versions of direct communication to citizens to measure the effect of information on switching. For example, a random subset could receive a letter at the end of the year from the Danish Energy Agency reminding them of the possibility to choose contracts freely and how to choose trustworthy suppliers. Using the smart meter data, one could then evaluate whether the information campaign was successful.

3.4. Voters will not like policy experiments

The basic premise of a policy experiment is that for some time a group of citizens or firms will be treated differently than another group. Moreover, the allocation to groups has to be done randomly. Generally, it is not unethical to treat firms or humans differently for the sake of understanding how to better design environmental policies. Medical science has a long tradition of conducting RCTs to determine the best possible treatments. If the outcomes of a policy or a treatment are certain, then yes, people should not be treated differently, but as long as it is unclear which policy is best, then it should rather be considered unethical to not try to find the best possible policy. Directing resources to inefficient programs can by highly unethical, as it means that environmental damages accumulate. Further, in medical trials the outcomes might be life and death while in policy experiments the outcomes are using less electricity or taking the metro more often. However, policy makers might still be afraid to suggest policy experiments because they are concerned about the reaction of voters. *Dur et al. (2023)* test four reasons why policy makers might be afraid to use policy experiments. First, the unfairness argument of treating individuals differently. Second, voters might feel that experimentation takes too much time and immediate action is needed. Third, voters might worry about the lack of informed consent when participants in an experiment do not know that they are part of the experiment, and forth, they might feel that experimental findings lack external validity. *Dur et al. (2023)* do a survey among ca. 2000 Dutch voters and among 126 Dutch politicians to study

their opinion on experiments. Generally, the find high approval among voters for policy experiments, with the support being highest when voters do not have a strong opinion about the policy. Less than 1 % of the participants always prefer implementation or no implementation over running an experiment first. Politicians also seem to be in favor of more experiments and strongly react to the information that voters are in favor of experiments. In a related paper on how consumers view corporate experiments, *Mislavsky et al. (2020)* summarize their findings as »Experiments are not unpopular; unpopular policies are unpopular.« Their participants only object to experiments when one of the tested policies is unpopular, not to the process of doing an experiment to understand what works. Based on this evidence, it seems that the fear of voter backlash against policy experiments is misplaced. In the Danish context, the experience with STAR and TrygFonden's Center for Child Research also shows that there seems to be little concern with experiments in general in the public.

3.5. That's not how we do things

A seldom discussed reason why policy experiments are not conducted in environmenta policy might be simply that nobody has started doing them. *DellaVigna et al. (2024)* call this »organizational inertia.« Governmental agencies have established ways of doing things and train new employees in these processes. While the questions might change, the method of working does not. Unless someone comes in and radically questions the way of working, the approach will not change. This radical change happened in the UK in 2010.

In 2010, David Cameron initiated the UK Behavioral Insights Team (BIT), also known as »The Nudge Unit.« The idea was to use insights from behavioral economics to inform all areas of policy. While the importance of using behavioral insights for policymaking deserves its own paper, an important element of the BIT's work was that everything they did was evaluated using randomized controlled trials. The need for RCTs was embedded in the unit, as they had a sunset clause: unless they brought in ten times the revenue they cost, they would be shut down after two years. One of the first experiments the BIT conducted became the most well-known policy experiment ever done *(Hallsworth et al. 2017)*. They collaborated with the tax authorities to improve tax collection in the United Kingdom. This experiment involved modifying the reminder letters sent to delinquent taxpayers. While some taxpayers received the standard letter, others received an additional sentence stating, »9 out of 10 people in your town/area/the UK pay their taxes on time. You are currently in the small minority that has not paid.« The additional statement led to an increase of £4.9 million (approximately $6.5 million) in (earlier) tax payments from a sample of almost 120,000 taxpayers. Because of its simplicity and impressive impact, this experiment has been widely discussed when promoting the benefits of behavioral science in public policy. The experiment alone justified keeping the BIT open, which has since turned into a private com-

pany with hundreds of employees. Since 2012, the effectiveness of tax reminders has been tested using RCTs in numerous countries, including Australia, Argentina, Austria, Chile, Costa Rica, Denmark, Germany, Guatemala, Israel, Peru, Switzerland, the United States, and many more. Most recently, *Holz et al. (2023)* partnered with the IRS of the Dominican Republic and found that a similar nudge increased tax payments by $193 million (0.23 % of GDP).

Even earlier, in 2008, a small start-up, O-Power, influenced by Robert Cialdini's work on social persuasion, conducted a trial in California comparing households' energy use to their more efficient neighbors. They added information on how the household compares to their neighbors' consumption and a happy or sad smiley on households' energy bills and found that this social comparison reduced electricity consumption by 2 %. In 2011, Hunt Allcott published a paper on their collection of trials with, at that point, 600,000 households (*Allcott 2011*). The paper has been cited close to 4,000 times, and home energy reports with social comparisons have been replicated hundreds of times by utilities worldwide with varying success. O-Power estimates that their initiatives have helped abate over 450,000 tonnes of CO_2 emissions and saved $75 million in energy costs.

Both of these interventions are excellent examples of how a small change in an existing process can have a large monetary impact when scaled up to thousands of people. They are fairly simple to replicate, inexpensive to implement, and, importantly, easy to measure. It is not surprising that they have been copied worldwide. Without a randomized controlled trial, they would not have had the impact they did. Imagine if the BIT or O-Power had added this additional information to the letters without an RCT. Would anyone have copied it without evidence that it makes such a difference? Most likely not. Without the use of RCTs for everything they do, the BIT would likely have been shut down after two years, and O-Power would not have been acquired by Oracle for $500 million in 2018 (*Staff 2016*). Similarly, many large tech companies such as Amazon, Netflix, Spotify, and Duolingo have experiments embedded in their company DNA, which undoubtedly contributes to their success (*Netflix Research n.d.*).

Hjort et al. (2021) even test the importance of evidence from RCTs for policy making directly in a field experiment. They randomly invited a subset of 1,818 Brazilian mayors to an information session explaining the impact of the behaviorally informed tax reminder letters and the benefits of conducting randomized controlled trials. 37.9 % of the invited mayors attended the information session. They then measured the implementation of this policy 15-24 months later and found that the invitation to the information session increased implementation by 33 % or 10 percentage points compared to the control group that was not invited. This provides clear evidence that experiments were adopted by those who were randomly invited to learn about a new approach to policy making.

3.6. We know what we are doing

Finally, policymakers might worry that those who favor experiments could be viewed as less competent because suggesting an experiment inherently admits that one does not know the answer to a question. As David Halpern, one of the founders of the BIT, writes in his behind-the-scenes book: »in order to get ahead in a political or management career it is better to be decisive and wrong, than uncertain and right« (*Halpern 2016*). While this approach might be ideal for individual careers, it is problematic if we want better policy results.

Being willing to run policy experiments requires two personality traits: admitting that one does not know the answer to a question and being willing to accept that policies implemented in the past might have been wrong. Even in medicine, the idea of rigorous testing is less than 100 years old, and there was significant resistance to testing established medical practices even in the 1960s and 1970s because doctors »knew what worked« (*Higgins et al. 2008*). However, today, no drug or treatment enters the market without being tested in a randomized controlled trial, and no doctor or medical scientist is considered less competent when they rely on causal evidence to suggest a treatment. There is nothing fundamentally different about testing medical treatments and policy treatments for their impact. In most cases, policy trials are easier and cheaper to run and have a larger impact on society.

From my own experience, many field experiments in both the private and public sectors have been stopped by someone in the organization who feared that the policy already implemented or the investments made would prove ineffective. The public transport agency that withdrew because the two people in charge of their engagement campaigns were not interested in knowing whether the campaign they had been running for years, costing thousands of dollars, was actually leading to more people taking public transport. The large stadium that did not want to know whether the climate labels they had designed and implemented actually changed food choices because then they might not be able to run the social media campaign about their latest sustainability measure. The supermarket chain that was unwilling to test whether their 2-for-1 campaigns increased food waste because the outcome could conflict with their profit goals. If there is no strong evidence for or against a policy, it is much easier to argue for whatever policy aligns best with political preferences.

Clearly, experimentation needs to be conducted with as much scientific rigor as possible and with transparency towards citizens. As *Wang & Yang (2021)* warn in their paper on policy experimentation in China, experiments must be conducted scientifically sound and without particular policy goals in mind. The goal of experimentation should be to answer questions, not to create evidence in favor of a particular agenda. Nevertheless, processes can be put in place to ensure fair evaluations, such as setting up ethical review boards that review experimental designs before they are conducted, and the requirement to publish public reports on experiments that have been done.

4. Providing funding for policy experiments

In addition to policymakers running experiments, there is enormous potential to learn from experiments conducted in collaboration with or by private actors. By providing funding conditional on creating causal evidence, policymakers can influence the quality of evidence they receive. For example, in 2020, the Swedish Food Agency opened a call to understand how policymaking could reduce household food waste. One of the requirements for the project was that the research method had to be a randomized controlled trial to receive funding. Together with Rambøll Sweden, I conducted a large-scale field experiment in eight supermarkets across Sweden (*Gravert et al. 2021*). We were interested in understanding whether multi-buy offers such as »buy 1 get 1 free« would lead to overpurchasing of fresh produce. We randomized the stores into four different treatments: a single discount, a multibuy discount, a version that made the reference price more salient, and one with a nudge that playfully reminded consumers not to purchase more than they needed. Over the two-week period, the stores displayed different discounts but kept the prices for two cucumbers/broccoli identical, regardless of treatment, to avoid any price effects. This means the cucumber under the single discount was priced at X/2 of two cucumbers under the multibuy treatment. The two other treatments had the same multibuy offer but differed in their framing. We then measured the amount of cucumbers/broccoli purchased under each discount and in each store. Because the offers were randomized and not announced in ads, we can be sure that the customers entering the stores are, on average, the same and that we measured the causal effect of the offer on purchase numbers. We found that consumers purchased 18 % more cucumbers when confronted with the »buy two for X SEK and save Y SEK« treatment compared to the »buy one for X/2 SEK« offer. Making the comparison price more salient and using the nudge also reduced purchasing compared to the multibuy offer by 11 % and 9 %, respectively. Using a household survey, we found evidence that the multibuy treatment also increased household food waste compared to the single discount treatment. The clear policy recommendation from the research is to restrict the use of multibuy offers for quickly perishable vegetables. While this has been discussed, it has not yet been legislated. The research findings have been presented at many policy forums within Sweden, in the UK, and even at a European Commission event, given the credibility of the results and their direct application to policymaking. By directing funding to experimental approaches, the Swedish Food Agency did not need its own capabilities to conduct an experiment but could ensure that the evidence they received was of the highest standard.

Similarly, the Swedish Environmental Protection Agency funded a three-year project at the University of Gothenburg for running field experiments on testing green nudges. My post-doc from 2015-2018 was funded by this project. During that time, several field experiments on green food choices were conducted by the team, such as a study on menu order effects at university canteens (*Kurz 2018*) and in a private restaurant (*Gravert & Kurz 2021*). In the latter study, we were interested in understanding whether placing vegetarian food at the top of the menu would change what restaurant patrons order. For three weeks in May 2016, the restaurant randomized who received which version of the menu—the one with the meat dish at the top or the one with the vegetarian dish at the top. We found that those who were randomly exposed to the vegetarian dish at the top ordered 25 % more vegetarian dishes than the other group and decreased their meat consumption without any negative effect on sales or customer satisfaction. The experiment was very simple to run. All the restaurant needed to do was print two versions of the weekly menu, which they already printed every week, and record which tables ordered which dishes, which they also already recorded. There were no additional costs. Since then, the effect of changing the menu order on food choices has been replicated many times and is a robust finding in the literature. It has also been shown to be far more effective than climate labels in affecting food choices (*Lohmann et al. 2024*). Policymakers could mandate that public kitchens present meal options in ascending order based on their climate footprint. There are dozens of food and health regulations that restaurants already need to adhere to, so mandating the order of how meals are displayed to reduce the climate impact would not be out of the ordinary.

5. Where should we start?

There will be questions that are more or less suitable for policy experiments. The question of whether banning red meat in public canteens will decrease their climate score does not need to be tested, as it mechanically follows from such a ban. The question of whether an industry-wide carbon tax will lead to leakage effects in that industry is impossible for one country to test and would be a question for the EU government to explore. Nevertheless, there are hundreds of policies that could be tested in this way.

Mainly, we want to focus on policies where the human factor could lead to a stark deviation from economic or engineering estimates. Often, we already have observational and anecdotal evidence that there is a gap between what economic or engineering models would predict and what we see in the field. In the area of energy use, this difference is known as the energy-efficiency gap. In 2008, a McKinsey study estimated that the US could save $900 billion in energy savings by 2020 through energy efficiency measures, and it has been heavily referenced

by policymakers (*Farrell & Remes 2008*). Similarly, the European Commission's »Energy Efficiency in Buildings« report estimates that the residential sector could achieve energy savings of around 30 % with the right investments and behavior change. To achieve these savings, however, people would need to make the required investments and change their behavior. Unfortunately, the evidence we have so far suggests that the engineering estimates are far too optimistic. For example, smart thermostats are generally considered to have enormous potential for saving energy if widely adopted. In 2017, the International Energy Agency estimated that they alone would reduce energy consumption by 10 % (*International Energy Agency 2017*). However, in a large-scale field experiment with millions of energy use observations, *Brandon et al. (2022)* show that smart thermostats actually increase electricity and gas consumption by 2.3 % and 4.2 %, respectively, because users frequently override permanently scheduled temperature setpoints. Similarly, *Alpizar et al. (2024)* find in a randomized controlled trial in Costa Rica that the actual impact of water-reducing technology is three times smaller than conventional engineering estimates (9 % vs. 28 %), and *Christensen et al. (2023)* conclude that home improvement simulations greatly overestimate the effect of energy savings through energy efficiency renovations.

Not only can engineering models be wrong, but economic models used to predict behavior have also been proven wrong numerous times. One of the most prominent economic theories is that providing monetary incentives will increase the desired behavior. A large share of policies use this assumption to steer behavior. In a recent study, *Jilke et al. (2024)* tested whether a monetary incentive would affect contributions to a public good. Moreover, they asked policymakers what they would expect the results of a monetary incentive on behavior to be, thus testing the reliance on economic theory. They asked 815 county heads, mayors, and municipal government representatives of towns in Germany with over 30,000 inhabitants to predict the effects of a financial incentive on COVID-19 vaccination. On average, the policymakers believed that the financial incentive would increase vaccinations by 15.3 percentage points. The researchers then tested the exact same incentive in a field experiment involving all 41,548 inhabitants of the German town of Ravensburg. The field experiment showed a precise null effect of the incentive. Studies evaluating the effects of monetary incentives on COVID-19 vaccinations have also shown little evidence of their effectiveness, yet incentives like these were widely implemented during the pandemic, wasting resources on ineffective policies (*Thirumurthy et al. 2022*).

It is plausible that the level of the implemented incentives in the mentioned studies was not suitable to change behavior. As the discussion on the price of carbon shows (*Drupp et al. 2023*), the question isn't whether there should be a price on carbon or not, but how high it should be to have the intended effects. This is where experimentation can play an important role. Rather than discussing and »arbitrarily« agreeing on a price, different levels could be tested and the most ef-

fective price chosen. In my study with Skånetrafik, the benefit of providing a one-month travel card compared to a two-week travel card far outweighed the additional costs. Here, providing a higher economic incentive made a difference in behavior. However, in my experiment on consumer switching in electricity markets, I found that the size of the incentives to switch was not relevant for decisionmaking. Whether consumers expected to save 500 DKK or 1000 DKK had no influence on whether they switched providers. In this case, other barriers were more important. This is important for policymakers to know, as we might otherwise assume that providing higher savings or subsidies would change behavior. In other situations, the barrier to behavior change isn't financial at all. In an early study by the BIT with the Department of Energy and Climate Change, they compared sending homeowners a subsidy to have their roof insulated at a low price to sending them a coupon for a junk removal company and loft insulation at a higher joint price. It turns out that, while financially less valuable, the group that received the junk removal coupon was more likely to show interest in insulating their roofs because the intervention addressed the right barrier—to insulate your roof, you first need to spend a weekend cleaning up your attic (*Behavioural Insights Team 2013*).

Lastly, while gathering knowledge about what works is important, it ultimately needs to lead to the implementation of effective policy. Therefore, we need to consider the likelihood of adoption by public authorities when choosing what to experiment on. *DellaVigna et al. (2024)* investigate what drives the adoption of experimentally evaluated nudges, like those discussed in this paper. They study 30 US cities that ran 73 RCTs with a national nudge unit. In 27 % of the cases, the cities adopted a nudge into their regular communication. They find that the strongest predictor of whether the nudge was adopted was not the effect size, but whether the communication to which it was added was pre-existing or new. If the nudge was embedded in pre-existing communication, the adoption rate was 67 %, while it was 12 % if the communication was new. The difference is large and highly significant. *DellaVigna et al. (2024)* write that the main problem with both experimentation and implementing new policies is mainly organizational inertia, as discussed in section 3.5. If there is no pre-existing process, then neither the experimentation nor, in their case, the tested communication is continued. This explains the huge success of the tax reminder nudges and the home energy reports, as every tax agency in the world will send reminders to their citizens to pay their taxes, and every utility will send bills.

6. Conclusion

Environmental policy has a lot to gain from adding policy experiments to their toolbox. Whether in collaboration with researchers, or independently, the experience of agencies such as STAR in Denmark show that policy experiments are a feasible and meaningful endeavor. A host of studies on environmental policy questions show that what we think should work, often does not and that there are multiple ways of how policy can be improved upon with evidence. Researchers are also limited in the policy-relevant work they can do without the collaboration of public agencies and companies. Collaboration would be a win-win for both sides. In this paper, I argue that the barriers to experimentation can be overcome if there is a will to challenge the status quo of how environmental policy is done. Given the ambitious climate agenda the Danish government and many governments around the world have set forth, we need to think about innovation not just in terms of technology, but also in policy making. What this article shows is that policy makers don't need to know all the answers. Instead, they can build in experimentation in the policy process and find out what works.

References

Allcott, H. (2011), 'Social norms and energy conservation', *Journal of Public Economics* **95**(9-10), 1082-1095.

Alpizar, F., Bernedo Del Carpio, M. & Ferraro, P. J. (2024), 'Input efficiency as a solution to externalities and resource scarcity: A randomized controlled trial', *Journal of the Association of Environmental and Resource Economists* **11**(1), 171-211.

Behavioural Insights Team (2013), 'Removing the hassle factor associated with loft insu- 655 lation: Results of a behavioural trial', *UK Department of Energy & Climate Change*, September. Available at: https://www. gov. uk/government/publications/loft-clearance-results-of-a-behavioural-trial.

Brandon, A., Clapp, C. M., List, J. A., Metcalfe, R. D. & Price, M. (2022), The human perils of scaling smart technologies: Evidence from field experiments, Technical report, National Bureau of Economic Research.

Christensen, P., Francisco, P., Myers, E. & Souza, M. (2023), 'Decomposing the wedge between projected and realized returns in energy efficiency programs', *Review of Economics and Statistics* **105**(4), 798-817.

De Quidt, J., Haushofer, J. & Roth, C. (2018), 'Measuring and bounding experimenter demand', *American Economic Review* **108**(11), 3266-3302.

DellaVigna, S., Kim, W. & Linos, E. (2024), 'Bottlenecks for evidence adoption', *Journal of Political Economy* **132**(8), 000-000.

Drupp, M. A., Nesje, F. & Schmidt, R. C. (2023), 'Pricing carbon: Evidence from expert recommendations', *American Economic Journal: Economic Policy*.

Dur, R., Non, A., Prottung, P. & Ricci, B. (2023), Who's afraid of policy experiments?, Technical report, Tinbergen Institute Discussion Paper.

Farrell, D. & Remes, J. K. (2008), 'How the world should invest in energy efficiency', *The McKinsey Quarterly* **11**, 1-I0.

Ferraro, P. J., Cherry, T. L., Shogren, J. F., Vossler, C. A., Cason, T. N., Flint, H. B., Hochard, J. P., Johansson-Stenman, O., Martinsson, P., Murphy, J. J. et al. (2023), 'Create a culture of experiments in environmental programs', *Science* **381**(6659), 735-737.

Frank, L. (2000), 'When an entire country is a cohort', *Science* **287**(5462), 2398-2399.

Gravert, C. (2024), From intent to inertia: Experimental evidence from the retail electricity market, Technical report, CEBI Working Paper Series.

Gravert, C. & Carlsson, F. (2019), 'Nudge som miljöekonomiskt styrmedel-att designa och utvärdera'.

Gravert, C. & Collentine, L. O. (2021), 'When nudges aren't enough: Norms, incentives and habit formation in public transport usage', *Journal of Economic Behavior & Organization* **190**, 1-14.

Gravert, C., Gunnarsson, E., Järneteg, A. & Leandersson, C. (2021), 'Kan insatser i butiken minska konsumenternas matsvinn?'.

Gravert, C. & Kurz, V. (2021), 'Nudging à la carte: a field experiment on climate-friendly food choice', *Behavioural Public Policy* 5(3), 378-395.

Hallsworth, M., List, J. A., Metcalfe, R. D. & Vlaev, I. (2017), 'The behavioralist as tax collector: Using natural field experiments to enhance tax compliance', *Journal of Public Economics* **148**, 14-31.

Halpern, D. (2016), *Inside the nudge unit: How small changes can make a big difference*, Random House.

Harrison, G. W. & List, J. A. (2004), 'Field experiments', *Journal of Economic Literature* **42**(4), 1009-1055.

Higgins, J. P., Green, S. et al. (2008), 'Cochrane handbook for systematic reviews of interventions'.

Hjort, J., Moreira, D., Rao, G. & Santini, J. F. (2021), 'How research affects policy: Experimental evidence from 2,150 brazilian municipalities', *American Economic Review* **111**(5), 1442-1480.

Holz, J. E., List, J. A., Zentner, A., Cardoza, M. & Zentner, J. E. (2023), 'The $100 million nudge: Increasing tax compliance of firms using a natural field experiment', *Journal of Public Economics* **218**, 104779.

International Energy Agency (2017), 'Digitalization set to transform global energy system with profound implications for all energy actors', https://www.iea.org/news/digitalization-set-to-transform-global-energy-system-with-profound-implic70a8tions-for Accessed: 2024-08-05.

Jensen, P. & Sjö, N. M. (2024), 'The effects of a large-scale school readiness intervention on danish preschool children's emergent mathematics skills', *Scandinavian Journal of Educational Research* **68**(3), 488-503.

Jilke, S., Keppeler, F., Ternovski, J., Vogel, D. & Yoeli, E. (2024), 'Policy makers believe money motivates more than it does', *Scientific Reports* **14**(1), 1901.

Kreiner, C. T. & Svarer, M. (2022), 'Danish flexicurity: Rights and duties', *Journal of Economic Perspectives* **36**(4), 81-102.

Kurz, V. (2018), 'Nudging to reduce meat consumption: Immediate and persistent effects of an intervention at a university restaurant', *Journal of Environmental Economics and Management* **90**, 317-341.

List, J. A., Sadoff, S. & Wagner, M. (2011), 'So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design', *Experimental Economics* **14**, 439-457.

Lohmann, P. M., Gsottbauer, E., Farrington, J., Human, S. & Reisch, L. A. (2024), 'An online randomised controlled trial of price and non-price interventions to promote sustainable food choices on food delivery platforms', *Available at SSRN 4818247*.

Mislavsky, R., Dietvorst, B. & Simonsohn, U. (2020), 'Critical condition: People don't dislike a corporate experiment more than they dislike its worst condition', *Marketing Science* **39**(6), 1092-1104.

Netflix Research (n.d.), 'Experimentation and causal inference', https://research.netflix.com/research-area/experimentation-and-causal-inference. Accessed: 2024-08-06.

Staff, W.P. (2016), 'Oracle agrees to buy arlington energy data firm opower for $532 million', https://www.washingtonpost.com/business/economy/oracle-agrees-to-buy-arlington-energy-data-firm-opower-for-532 million/2016/05/02/83739416-107f-11e6-93ae-50921721165d_story.html. Accessed: 2024-08-05.

Tagesschau (2024), 'Penny-aktion: Lebensmittelkosten inklusive der umweltfolgekosten', https://www.tagesschau.de/wirtschaft/verbraucher/penny-umwelt-folgekosten-102.html. Accessed: 2024-08-01.

Thirumurthy, H., Milkman, K. L., Volpp, K. G., Buttenheim, A. M. & Pope, D. G. (2022), 'Association between statewide financial incentive programs and covid-19 vaccination rates', *PLoS One* **17**(3), e0263425.

Wang, S. & Yang, D. Y. (2021), Policy experimentation in china: The political economy of policy learning, Technical report, National Bureau of Economic Research.