Experimental innovation policy

Albert Bravo-Biosca[1]

Innovation Growth Lab - Nesta and
Barcelona Graduate School of Economics

DRAFT

2 April 2019

Abstract

Experimental approaches are increasingly being adopted across many policy fields, but innovation policy has been lagging. This paper reviews the case for policy experimentation in this field, describes the different types of experiments that can be undertaken, discusses some of the unique challenges to the use of experimental approaches in innovation policy, and summarizes some of the emerging lessons, with a focus on randomized experiments. The paper concludes describing how at the Innovation Growth Lab we've been working with governments across the OECD to help them overcome the barriers to policy experimentation in order to make their policies more impactful.

[1] This paper has been written as a contribution to the NBER Innovation Policy & Economy Series. The author gratefully acknowledges the financial support received from the National Bureau of Economic Research (NBER) and the partner organizations of the Innovation Growth Lab. This paper builds on prior work undertaken by the author with other colleagues at the Innovation Growth Lab, in particular Teo Firpo, James Phipps and Lou-Davina Stouffs. The author would like to thank them for their contribution to this paper and also the editors for their comments. All errors are my own. Correspondence address: Nesta, 58 Victoria Embankment, London EC4Y 0DS, United Kingdom. Email: abravobiosca@nesta.org.uk.

1. Introduction

The main aim of innovation policy is to support experimentation with new technologies, products, processes, or business models, and accelerate its diffusion throughout the economy and society. Yet innovation policy per se is not very experimental. Policymakers invest billions funding many scientific and business experiments, but they rarely experiment themselves with their own programs and activities, at least in a structured way.

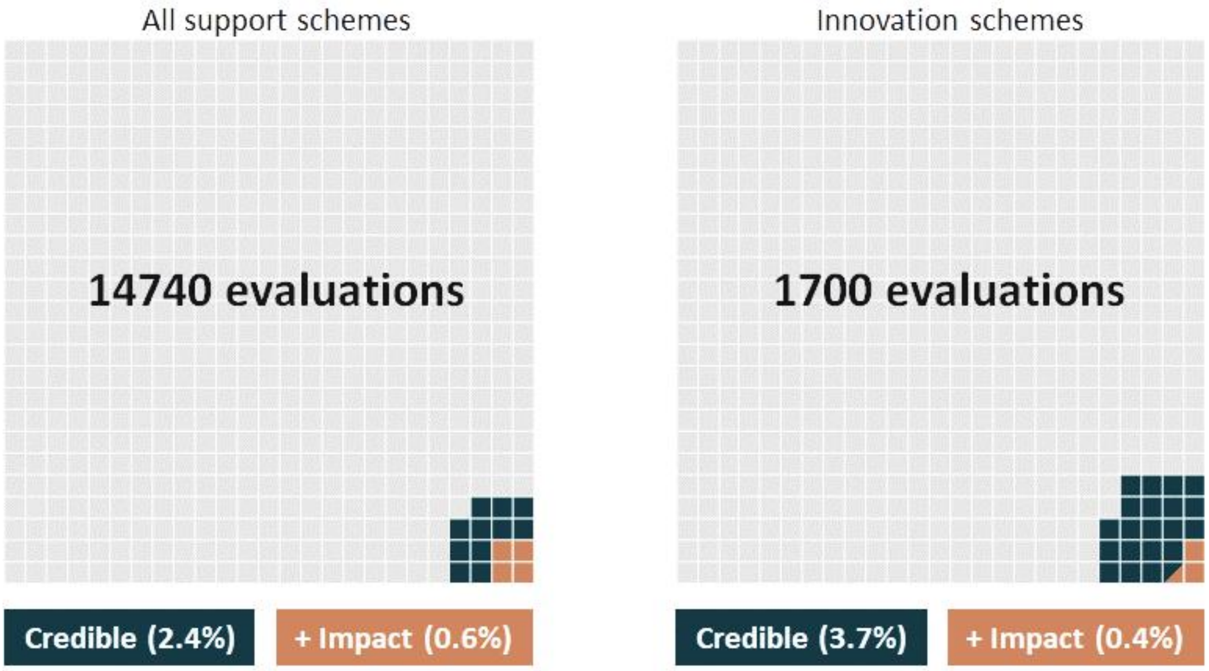
Are we making the most of this investment? Are there more effective ways of using this funding? How would we ever know? These are questions that we need to address if we want to successfully navigate the economic challenges we face ahead. Yet in many different ways we are in uncharted territory, for at least three reasons.

Firstly, innovation systems are difficult maps to chart. They are complex systems rather than simple linear production functions. Actors, institutions and policies interact in multiple ways, and levels of uncertainty are high. Shifting a policy lever may have unanticipated consequences due to priory unknown interdependences, so making predictions and allocating funding are not easy exercises. It takes time to shed some light on how a system works.

Secondly, innovation systems are continuously evolving, some argue that faster than in the past. Some of the trends reshaping innovation systems include the raise of global value chains, the globalization of knowledge production beyond OECD countries, the increasing burden of knowledge (Jones, 2009), new general purpose technologies (such as AI and digitization), increasing market concentration and changing dynamics between startups and corporates. In a changing context, old solutions may not work (if they ever did). Many of these also give raise to new challenges that have not been encountered before, and which will require imaginative solutions (such as climate change or the transformation of work). In parallel, emerging technologies may also offer new and unexploited opportunities for policymakers, although it is unclear how best to take advantage of them. For instance, how will AI change innovation and innovation policy (Cockburn et al, 2018)?

The last reason why we are in uncharted territory is that we lack much of the evidence that we would need to guide policy decisions. Some years ago the UK foundation Nesta funded the University of Manchester to develop the Compendium of Evidence on the Effectiveness of Innovation Policy (Edler et al, 2016). It was full of insights, but it was also somewhat discouraging. Many policy areas had little evidence, others had very poor quality evidence, and for policies that had good evidence, it often showed that their effects were small or negligible.

Figure 1: How good is the existing evidence base? Robustness level of existing impact evaluations



Source: Charts based on the systematic reviews conducted by the LSE-based What Works Centre for Local Economic Growth (Credible: Level 3 Maryland Scale – Positive impact on employment)

More recently, the What Works Centre for Local Economic Growth at the LSE examined almost 15,000 impact evaluations of local economic policies, assessing the methodology that they used and their results.[2] As seen in Figure 1, they concluded that only 2.5 percent of them had a credible counterfactual and provided strong evidence of causality.[3] Most of the other evaluations, while still containing useful insights, were not rigorous enough to be able to convince someone who disagreed with the conclusions to change his mind. In other words, they provided suggestive correlations, rather than strong evidence that the program being evaluated had (or had not) caused a change in the outcomes. The review also found that, among the "credible" evaluations, only one in four demonstrated a positive effect on employment (or 0.6 per cent of the total).

[2] All Policy Reviews available at www.whatworksgrowth.org/policy-reviews.

[3] "Credible' refers to evaluations that satisfy the level 3 of the Scientific Maryland Scale, which requires that the evaluation method used has a credible counterfactual. Note that random allocation is not a requirement for level 3, it is sufficient to have a clear justification for why the control group would have performed in a similar way as those benefiting from the intervention if the intervention had not happened.

This is not to say that we should aspire all innovation policies and programs to reach the highest standards of evidence. Many important questions cannot really be answered without doubt. If all evaluations provided incontrovertible evidence of causality, it would mean that we are not asking many of the most relevant questions. But there are still many questions for which causal inference would be feasible and useful. There is definitely immense scope to increase the quality and quantity of evidence in this policy space (although it is important to ensure that the resulting evidence is both useful and used for the investment to be justified).

In short, innovation policymakers face a complex and continuously evolving system and have very limited evidence on how most effectively to influence it. The question is how can we start to navigate all the unknowns and shed some light on the possible answers. One alternative is to become more experimental. That is, exploring a wide range of ideas, testing out the most promising ones at small scale, learning which are likely to work better, and only then scaling them up.

This would mean turning the current model of policymaking upside down. Despite all the unknowns, governments often act as if they had all the answers, rather than recognizing that they do not. They introduce new policies without prior small-scale testing, assuming they have chosen the best design and hoping it will work.

Would other approaches have achieved more impact, or been equally successful while using fewer resources? Which design of the program - the devil is often in the details - would be most effective? Questions such as these are often left unanswered, as public agencies struggle to fit political priorities in a short policy cycle. Ultimately, this leads to less effective policies and the risk of wasting limited resources on programs that don’t work.

The UK provides an interesting example of how policymakers can embrace a more experimental approach. The business ministry (BEIS) wanted to encourage small businesses to seek external advice in a range of areas, from digital technologies to management skills. It launched the “Growth Vouchers” program, a $40 million pilot which gave small businesses vouchers of up $2,500 to use in a marketplace of business providers. Rather than starting with a single policy design, the whole program was conceived as a policy experiment. Within it, there were a number of randomized trials, not only testing the impact of the program but also different modes of delivery (e.g., from different messages to attract applicants to different tools for the assessment phase). At the IGL2016 conference, the senior civil servant leading the program was asked by one of the attendees in the room what if the program was shown not to work. His answer was clear: "*We will have saved a lot of money*". Without evaluation, policies that don't work may continue indefinitely, depriving resources from more impactful interventions.

In this paper we describe why an experimental approach can contribute to more effective innovation policies, how policymakers can become more experimental, and the work that we have been doing at the Innovation Growth Lab (IGL) to help them in this process.

IGL was set up in 2014 by the UK foundation Nesta and the Kauffman Foundation in the US. It is a global partnership that brings together governments, foundations and researchers to test different approaches to accelerate innovation, entrepreneurship and growth. Our shared ambition is to make innovation and growth policy more impactful though experimentation and evidence.

This paper is structured as follows. The next section discusses what does it mean be of experimental. Section 3 focuses on a particular type of policy experiments, randomized controlled trials, and why, when and how they can be used. Section 4 summarizes some of the evidence emerging from randomized experiments in innovation policy. Section 5 addresses the barriers to innovation policy experimentation, and section 6 concludes.

2. What does it mean to be experimental?

The word "experiment" is often used in many different ways, so it is useful to clarify what the meaning of an experiment actually is. In short, an experiment is a test. More specifically, the Cambridge English Dictionary defines experiment as "*a test done in order to learn something or to discover if something works or is true*".

This definition captures the key characteristic of a policy experiment: learning. It is intentionally set up to learn. It has a clearly structured learning strategy, defined ex-ante rather than as an after-thought, and generates new information, evidence or data. Therefore, a government pilot "trying something new" is not a policy experiment, unless the systems and processes required to learn from it are also put in place. This includes a timeframe with clear limits or checkpoints: there is date at which you assess the results and decide whether to continue the experiment, tweak it, scale it up, or discontinue it.

Ideally, policy experiments start at a small scale, not being larger than what is required to answer the question or validate the hypothesis being tested. Whenever feasible and appropriate, they have some form of control group, but this is not a pre-requisite (although having one makes learning much easier). Lastly, it is good practice to codify the knowledge created by the experiment, so that it can be shared, replicated, and built upon it.

This definition of an experiment is both wide and narrow. Wide because it tries to capture a range of experimental approaches that are used in different disciplines, from design to economics. But narrow because it does not include unintentional or natural experiments. These are not deliberately set up to test something and therefore learning is not a priority, but they still create retrospective learning opportunities that can be exploited using observational data. For example, when governments use lotteries as a low-cost mechanism to allocate participants in an oversubscribed program (Cornet et al, 2006); when geographic boundaries or bureaucratic processes create discontinuities that can be exploited using econometric methods (Criscuolo et al, 2019); or when a federal system creates opportunities for regions to use different policy tools to address similar challenges, which retroactively might be thought of parallel experiments and can analyzed with both quantitative and more qualitative approaches (Ansell and Bartenberger 2016).

Experiments are at the core of policy experimentation, but the process of experimentation involves other important steps. It starts with understanding the problem, creatively exploring unobvious ideas, and developing hypotheses and potential solutions that can be tested. It does not end when the results of the test become available. Instead, governments that have successfully embraced a culture of experimentation not only set up experiments, but they also make sure the resulting learning and evidence is used in decision-making, scaling-up successful ideas while continuing to iterate and experiment.

A (very) simple typology of policy experiments

Policy experiments can be used in different contexts and with different objectives. Table 1 tries to distinguish some broad types of experiments and their underlying motivation. They can be divided into two groups: those that are focused on exploration and discovery (understanding how the world works), and those framed around evaluation (finding out what works).

Within the first group, mechanism experiments can be used to test assumptions about the problem to be fixed, the underlying drivers of behaviors, or the solution being considered. Scientific experiments constitute the best example: scientists develop a theory, derive a set of hypotheses from it, and test them in order to prove or disprove the underlying theory. Policy experiments within this category have a similar ethos. Their main aim is not to understand whether a particular intervention works or doesn’t, but rather testing whether the mechanisms proposed by the theory or the assumptions that underlie it hold or not (“theory” in this context can refer to an economic theory modelling human or firm behavior, but also a theory of change for a specific program).

Alternatively, experiments can also be used to explore the feasibility and potential of a new intervention: Can it be delivered? What types of outcomes are likely to emerge? How do people or businesses respond to it? These exploratory experiments seek to answer the "what if" question (Ansell and Bartenberger, 2016), exploring expected and unexpected consequences rather than seeking conclusive answers. They can be very useful in situations in which there is high uncertainty and limited prior knowledge to build on, but their potential uses extend beyond that. They often involve setting up prototypes and continuously iterating and adapting their design in order to learn how to improve them through trial and error.

Table 1: Types of policy experiments

	Mechanism experiments	**Exploratory experiments**	**Optimization experiments**	**Evaluation experiments**
Main aim	**Exploration and discovery**		**Evaluation: what works?**	
What is tested	**Assumptions** Testing assumptions about the problem to be fixed, the underlying drivers or mechanisms of observed behaviors, or the solution being considered	**Potential** Testing the feasibility and potential of new solutions, exploring expected and unexpected consequences rather than seeking conclusive answers	**Process** Testing process changes (small or large) in order to optimize the process used to deliver an intervention (not looking at the impact on outcomes but rather on inputs and outputs)	**Impact** Testing the impact of an intervention on outcomes, or comparing the effectiveness of different interventions (or versions) in order to find out what works, when, and for whom
Learning method	Randomized controlled trials, rapid cycle testing, A/B testing, mixed methods, ethnographic research, human-centered design, prototyping, other qualitative & quantitative approaches			
Common characteristics	1. Learning is the priority: generates new information, evidence or data 2. Intentionally tests or trials a defined idea or hypothesis 3. Has a structure: a systematic process that allows learning to happen 4. Timeframes set from the start to assess results and make decisions			

The second group of policy experiments are focused on evaluation, although from two different perspectives: impact evaluations that estimate the ultimate impact of an intervention on outcomes, and process optimization experiments that measure intermediate impacts of changes in the process.

Impact evaluations are one of the most common type of experiments. They may be used to evaluate a single program, to test the impact of small tweaks in a program, or to compare the impact of two or more different programs. The key question they seek to answer is what works, when, and for whom. Consequently, they always try to measure the outcomes that policymakers are trying to influence.

Increasingly, it is becoming more common (and easier) to use experiments to optimize the processes used in the delivery of a program. These do not seek to measure how outcomes change, but rather improve one of the steps involved in the delivery of the program. The underlying assumption is that this optimization will result in more efficient and impactful programs, but this assumption is not actually tested. A common example are A/B experiments that test ways to increase the number of participants applying to take part in a program. Many of these experiments happen “under-the-radar”, embedded into day-to-day operations, and as a result the finding are often not codified.

These four categories of experiments are not mutually exclusive. For instance, some experiments may try to test a theory and an intervention simultaneously (asking what works and why it works), or use process optimization trials (like A/B testing) to test some theoretical mechanisms.

In some other cases, these different types of experiments may be undertaken sequentially: starting with a prototype first, following with a full-fledged impact evaluation, and finally refining the intervention testing tweaks in the process. Where to start ultimately depends on our prior knowledge. Do we know what outcomes to expect? Or we don't really know what is likely to happen as a result of the experiment? Do we have prior evidence about the potential of the solution, or we don't really know whether it is implementable yet?

To give a concrete example, imagine a big science lab that would like to encourage serendipitous interactions between research groups in order to increase interdisciplinary collaboration. A range of options may include a central coffee machine, weekly lab drinks, yearly research retreats, etc. You may ask first the “what if” question: what happens if we put a nice Nespresso coffee machine in the middle of the lab and then carefully observe the behaviors of researchers when they use it. Are there more informal interactions between researchers from different teams? Do they discuss research projects or last night's football game? Do researchers become more addicted to coffee? What about non-coffee drinkers?

If the intervention appears promising, you may ask whether it really works. For instance, adding coffee machines in a random set of floors, tracking whether there are more follow-up email conversations or meetings between researchers from different groups in floors with coffee machines, measuring whether these lead to new research collaborations, and estimating whether there are spillovers and non-coffee drinkers also benefit.

You may also consider how to optimize the process. Are more conversations initiated if the coffee machine is slower at preparing coffee, giving more time for interaction? Are the conversations more productive if there are tables and stools around the coffee machine? What is the optimal number of coffee machines, and where should they be positioned? Does sending coffee email reminders to random pairs of researchers makes it more likely that they will begin a conversation?[4] Does it make a difference if the coffee is free or needs to be paid for?

Lastly, it may also be possible to use the coffee machine experiment to test some assumptions about the problem you are trying to fix. For instance, how do within-lab networks get formed? Is the main barrier to interdisciplinary collaboration not knowing about each other's work, not having a personal connection with researchers outside your field, or is it a mismatch of interests and/or incentives?

The final question to consider when thinking about different types of experiments is how we learn from them. There is a range of methods that can be used, including randomized controlled trials (RCTs), A/B testing, rapid cycle testing, ethnographic research, human-centered design or mixed methods (among many other qualitative and quantitative approaches). Importantly, there is no one-to-one mapping between learning methods and the four types of experiments outlined in Table 1, so we need to avoid the temptation of allocating them this way. For instance, RCTs can be used to test assumptions, processes and impacts, as can ethnographic research. In some circumstances the only feasible method to evaluate an ecosystem level intervention may be a carefully conducted case-study. Typically, the approach that produces the most robust evidence of causality is mixed methods, combining both quantitative and qualitative approaches (rather than choosing between them, a false dichotomy). Ultimately, the choice of

[4] For instance, Nesta runs a randomized coffee meetups program that randomly pairs staff members for coffee breaks to learn about each other's work (See www.nesta.org.uk/blog/institutionalising-serendipity-via-productive-coffee-breaks).

method relies on the question being asked and the context in which the experiment is taking place, which determine what is feasible and desirable.

Experimentation in innovation policy

Innovation policy can itself be conceived as a continuous learning and discovery process (Bakhshi et al, 2011), about new technologies, the inner workings of the innovation system, and the effectiveness of programs and policies that seek to influence it.

The four types of experiments in Table 1 play a role in this process. As the above example illustrates, there are a wide range of experiments that can be conducted even when considering a very simple intervention (e.g., a coffee machine). Obviously, the toolkit of innovation policymakers is much wider (and impactful) than this example, exposing how many missed opportunities for experimentation actually exist.

Innovation experiments can be used to understand how different types of innovation processes or methods work (Boudreau and Lakhani, 2016), which in turn can also generate useful insights that inform the design of new programs and policies. Alternatively, they can also be used with a program evaluation mindset, to test whether a program works and how it can be improved. Finally, experiments can be framed around specific policy challenges, and be used to explore solutions that contribute to address them. As a result, experimentation on innovation policy not only happens in innovation agencies and ministries, but often also across other government departments addressing sectoral challenges (e.g., smart mobility labs).

The case for experimentation in innovation policy is reinforced by the complexity of the system that innovation policymakers try to influence, a very dynamic context that continuously evolves (with new challenges and opportunities regularly emerging), and high levels of uncertainty (in terms of policy levers and potential interactions, returns on investment from programs, or future scenarios among many others).

To confront this challenge, it is important that policymakers recognize that they do not have all the answers. Designing a new program or policy to support innovation involves a large number of decisions and choices. Many of these questions cannot be reliably answered by looking at past experiences or the literature, or by undertaking sophisticated foresight exercises. One alternative (the most common one) is to try to guess the best answers and proceed as if they were the right ones. However, it is more effective to test different answers to find out which one is the right one. And crucially to do this as the program is being designed and rolled out in order to obtain timely information, rather than not doing it at all or only doing it many years later when conducting an ex-post impact evaluation.

What would that mean in practice? First, making more use of design methods when developing new programs, as the Australian Department of Industry or the Polish Agency for Enterprise Development (PARP) have done by setting up in-house policy design labs (BizLab and InnoLab respectively). Second, developing new pilot programs explicitly as experiments. For instance, Sweden's national innovation agency (Vinnova) launched an experimental program that placed makerspaces within hospitals to increase user-led innovation within the health sector (Svensson 2017). Third, making more use of randomized trials (RCTs), which have been particularly underutilized in this policy space. This has been the main focus of IGL's work with governments, even if increasingly we have also been exploring other complementary approaches that also contribute to making policies more experimental and impactful.

The sections that follow focus on randomized trials, even if much of what is discussed is also valid for other experimental approaches. We describe why and when randomized trials are useful, what we are learning from them, and how at IGL we've been working with policymakers to help them overcome the barriers that limit their use.

3. Randomized trials in innovation policy

What are randomized trials, and why are they useful?

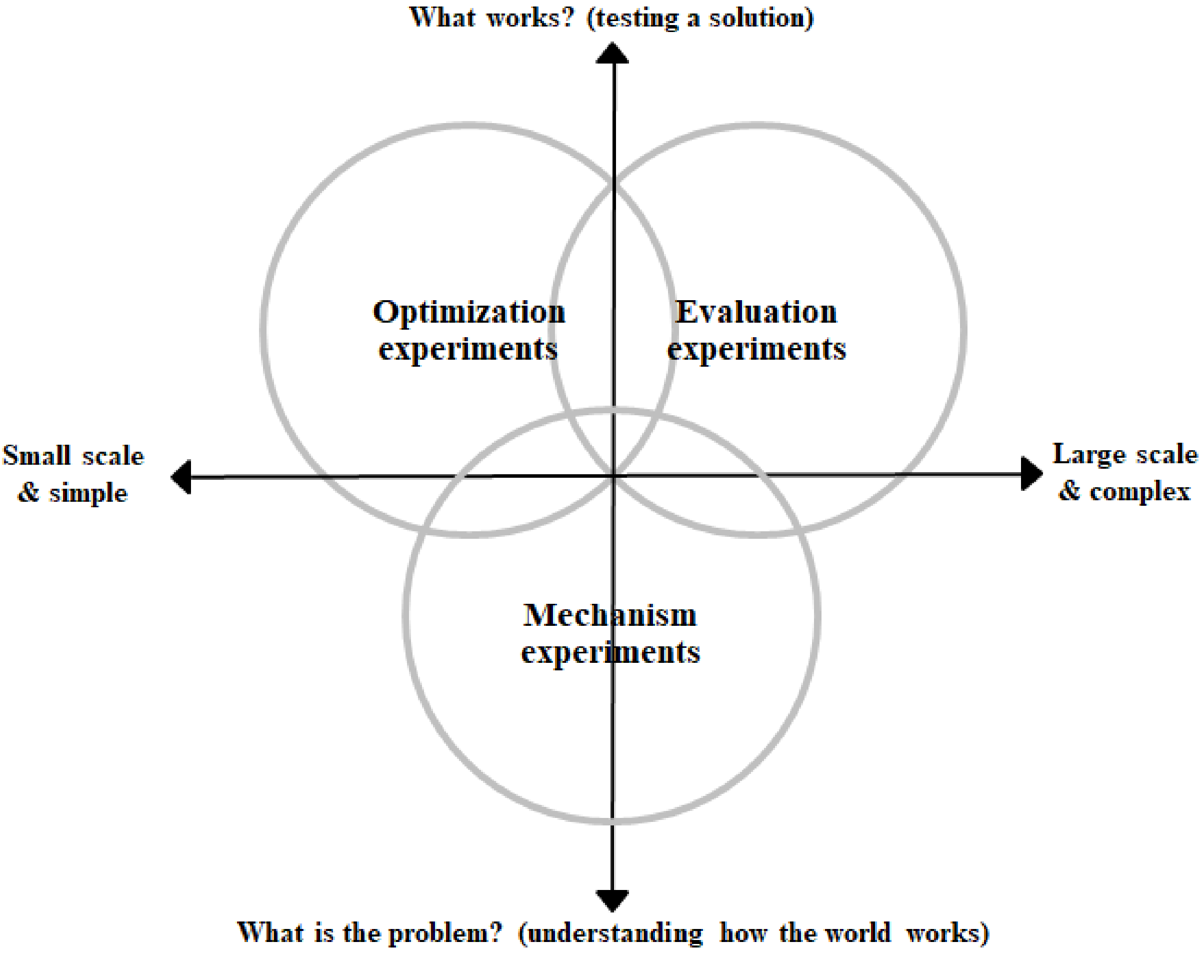
The central idea of randomized controlled trials (RCTs) is to allocate whatever is being tested by lottery. Specifically, participants are randomly placed across different groups, and the impact of the intervention(s) is estimated comparing behaviors and outcomes across them. The lottery used to assign participants to each group addresses potential selection biases. As a result, different groups are in principle comparable and any differences between the groups are the result of the intervention (as long as the sample size is sufficiently large). Therefore, randomized trials can provide an accurate estimate of the impact of a program.

In doing so, randomized trials address a common pitfall of public policy evaluations. Typical evaluations of innovation, entrepreneurship and small business programs only give a good answer to the question "how well did the program participants perform before and after the intervention?" They commonly fail to provide a compelling answer to the more important question: "what additional value did the program generate?" Or in other words, is the improved performance of firms receiving the intervention the result of the program itself, or does it reflect some unobserved characteristics of the firms that chose (or were selected) to participate in the program? Answering this question requires good knowledge of how participants would have performed in absence of the program, which is difficult to find out unless you have a credible control group that provides a counterfactual. Randomized trials achieve this by creating two truly comparable groups - only differentiated by the randomization process (the lottery). In contrast, many other evaluations fail to create a credible counterfactual. As a result, they are only convincing to those who are already predisposed to agree with the evaluation findings, but fail to convince those who have other views.

High-quality evaluations with a credible counterfactual are robust enough to change people's views on the impact of a particular program, and therefore are more likely to influence the choices that are made, leading to better decisions. They can also contribute to protect future investments in successful programs from changes in government and political priorities. Because of this, randomized trials are often referred to as the "gold standard" for evaluation, although as with any other method they have their uses and limitations (Deaton and Cartwright, 2018). There are other approaches that can also be used to identify a credible counterfactual from existing observational data and generate robust evidence. Therefore, the decision on which method to use depends on the characteristics of the program and the circumstances under which it is implemented. Mixed methods (combining quantitative and qualitative approaches) often provide the most insightful and robust answers, although they are not always feasible. As we discuss later in this section, some important questions cannot be addressed with counterfactual evaluation methods, so alternative approaches are also needed.

Randomized trials have been used extensively in health to test the effectiveness of new pharmaceutical drugs as well as medical procedures. But they have also been widely adopted in several other policy areas, such as development, education or social policy. For instance, the Abdul Latif Jameel Poverty Action Lab (J-PAL) at MIT has run over 900 randomized trials of poverty-reduction interventions in over 75 countries, and together with Innovations for Poverty Action (IPA) has radically transformed the development field in the process. The UK-based Education Endowment Foundation is conducting over 130 randomized trials involving more than a 1,000 schools and 900,000 pupils in order to test different ways to improve educational outcomes. And the French government runs an experimentation fund for young people, a bottom-up approach to identify innovative interventions to improve youth outcomes (crowdsourced from organizations across the country), implement them at a small scale, and rigorously evaluate them to find out whether they work, before deciding whether they should be scaled up.

Figure 2: Randomized trials in innovation policy



In contrast, the use of randomized trials to test innovation, entrepreneurship and small business programs has been very limited, particularly in advanced economies, despite frequent calls from the research community to increase their use (e.g., Azoulay 2012, Boudreau and Lakhani 2016). Among the different methods available in the evaluation toolkit, randomized trials have been particularly underutilized in this domain, and the quality of the evidence has suffered as a result. This is starting to change. IGL maintains an online repository of randomized trials related to innovation, entrepreneurship and business growth, describing each trial and summarizing their key results and policy implications.[5] At the last count the database shows a total of 130 trials, including both completed and on-going trials, with roughly half of those having taken place in the past six years.

[5] Available at www.innovationgrowthlab.org/igl-database.

Impact evaluation is one of the uses of randomized trials in innovation policy, but as discussed in Table 1 in the prior section, their potential use extends beyond that. As shown in Figure 2, they can also be used to test innovation theories and the underlying mechanisms that drive behaviors, as well as to optimize the processes used to deliver an intervention.

The innovation policy questions that randomized trials can(not) address

Running randomized trials on innovation policy questions can be more difficult than in other fields for several reasons. First, the outcomes of innovation policies are not always easy to measure. Innovation can be a "fuzzy" concept, and existing metrics of innovation are only incomplete proxies (from patents to high-tech startups). In contrast, outcomes tend to be much easier to measure in other fields in which randomized trials have been more widely used, such as health (e.g., survival or quality-adjusted life years), education (e.g., test scores) or development (e.g., income or poverty rates).

This is a common challenge for all evaluation methods in innovation policy, not only randomized trials. However, trials frontload the evaluator's work, so that the majority of the planning, decision-making, and analysis design happen before the intervention has even started, unlike in observational studies. This has its advantages, but it also means that once the trial has begun it is very difficult to change any of its parameters. Therefore, it is important to identify the right measures that capture the specific outcomes that the intervention seeks to influence, ideally using a detailed logic model or theory of change. For instance, an intervention might aim to improve collaborations between SMEs and universities. A simple measurement, such as number of collaborations, might miss a more profound change taking place as a result of the intervention (such as higher frequency of interactions or larger-scale/longer projects). Because the baseline survey can only be run once, asking the wrong question can compromise the whole project.

A second challenge is that outcomes can take longer to become visible than in other fields. Innovation is often a long process, and the channels through which innovation policies work can take a long time to impact observable outcomes. As a result, by the time the results of randomized trials become available they may be of little use, particularly if the policy no longer exists or it has been changed substantially (although historically innovation policies have evolved very slowly, and even today many policies are similar to their equivalents from decades ago). In order to get more timely results, it is useful to identify intermediate outcomes that become visible much earlier in the process, and which according to the theory of change of the program and existing empirical evidence predict changes in the ultimate outcomes (while in parallel putting in place the systems to track long term impacts).

Third, innovation outcomes can be very skewed. Most innovation projects fail, particularly if they are radical rather than incremental. Extreme successes are very rare, yet these are often the ones that many public policies are targeting (e.g., the "unicorns" or "blockbuster drugs"). Randomized trials work well to compare average performances, but require substantially larger samples to identify with statistical confidence impacts on the extremes of the distribution. If these are the impacts that policymakers would like to measure, then randomized trials might not be the most appropriate approach (and it might be better to rely on historical observational data that includes the universe of firms, even if it is more difficult to demonstrate causality).[6] However, similarly as above, an alternative is to identify intermediate outcomes that are less skewed, building on the program's theory of change and existing evidence (e.g., raising VC is an intermediate outcome that is positively correlated with becoming a unicorn). If the policy does not have an impact on intermediate outcomes, it is unlikely (even if not impossible) that it will impact ultimate outcomes. On the contrary, if the policy impacts intermediate outcomes positively, then it is more likely that ultimate outcomes are also improved.

Fourth, innovation ecosystems are complex environments, with observed and unobserved linkages and interactions, which make it more difficult to accurately predict the impact of a policy. Context and historical path-dependencies are particularly important, so even if randomized trials have high internal validity, external validity also needs to be assessed carefully to make sure the results can be useful and the investment in the trial is justified. Whenever possible, it is useful to test similar interventions in different contexts in order to understand when the results generalize, and when they don't. Similarly, learning should not end when the trial ends. If the decision is made to scale up a program, an evaluation should be set up alongside, since a small-scale program that works in a very specific context may not work as well at a larger scale across very different settings.

Lastly, many innovation policy challenges are multidimensional, and so is the solution space. In contrast, randomized trials are designed to test binary choices. A hypothesis can be true or false. While it is possible to test multiple hypotheses, or compare multiple treatments, there is a limit to the number of options that can be tested at once. In situations with limited prior knowledge and high uncertainty on the context, the intervention potential and/or the likely outcomes, it is useful to reduce the choice set and identify the most promising design through more exploratory experiments, prototyping and iterating through trial and error, prior to setting up a randomized trial. One exception to this is when the cost of setting up randomized trials is very low and outcome data is available almost immediately (such as in A/B

[6] This is particularly the case if the interventions being considered have skewed outcomes, are not very intensive and have relatively small effect sizes.

online experiments), in which case it might be possible to continuously test multiple binary choices and ultimately address multidimensional questions.

None of the challenges above are insurmountable. How easy it is to address them depends on the policy being considered and the aim of the experiment (i.e., impact evaluation, process optimization or mechanism experiments). In some cases, the compromises required may make the use of randomized trials unfeasible or undesirable, while in others trials can add substantial value.

The menu of innovation policies is wide. It includes programs that directly support innovators, entrepreneurs or businesses (such as entrepreneurship training, R&D grants, science funding, or tech transfer schemes), programs targeted at improving the functioning of the ecosystem (such as venture capital schemes or infrastructure), and a wider set of policies that set up the framework conditions (such as regulation and tax policy).

There are also differences in the underlying rationale for government intervention. Innovation policy can be framed around missions (such as climate change), system failures (such as missing actors and connections), or market failures (such as externalities). However, this does not impact how feasible and desirable it is to experiment. What ultimately matters is not the underlying rationale but rather the nature of the policy instrument and the questions being asked about it, as summarized in Table 2.

Table 2: Potential uses of randomized trials in innovation policy

	Mechanism experiments	**Optimization experiments**	**Evaluation experiments**
Framework conditions (e.g., tax, regulation)	Medium	Medium	Low
Ecosystem (e.g., clusters, infrastructure)	Medium	Medium	Low (overall) Medium (tools)
Targeted programs (e.g., grants, advice)	High	High	High

From an impact evaluation perspective, randomized trials can be used to evaluate policies and programs which have a targeted population that can be randomized into different groups, such as entrepreneurship and business support schemes among many others. It requires the ability to determine the treatment or intervention that participants in each group will be subject to, and also the possibility (ideally) to exclude participants from self-selecting into a particular group or intervention.[7] The trials may seek to estimate the impact of a program (comparing the outcomes for the treatment and the control group) or alternatively to compare the impact of two different versions of a program, without necessarily having a control group (often an existing program vs. a modified version that the organization is considering to introduce). For instance, when rolling out an innovation funding scheme for SMEs, an experiment can be used to test the impact of the scheme, but also to test whether adding a management coaching element on top of it makes the funding more effective.

On the contrary, it is not possible to use a randomized trial to evaluate the overall impact of an ecosystem or national-level policy intervention[8], or to select how to prioritize public investment between different missions, research fields, themes, or large infrastructure investments. However, many ecosystem-level policies consist of a bundle of instruments or activities, which might still be possible to evaluate with randomized trials.

For instance, the overall impact of cluster policy cannot be evaluated with a randomized trial (unless you are in the unlikely scenario of being able to randomly pick where new clusters are being set up). However, the delivery of cluster policies often includes a series of targeted programs, for which randomized trials are feasible evaluation approaches. While they will not provide the full impact of policy, since cluster interventions are intended to be more than the sum of its parts (with complementarities and interactions between instruments being a key element), they will still contribute to understand its impact. There is usually nothing preventing an organization from using several approaches to understand the effects of a policy - and in this sense randomized trials can be used as part of a larger evaluation strategy.

When the main motivation to experiment is not to evaluate a policy, but rather to test ways to optimize the processes used to deliver it, then the opportunities for experimentation are much larger. Process

[7] An exception are randomized encouragement designs, a type of experiment in which everyone is free to take part on a program or use a scheme, but only a random sample of potential participants are “encouraged” to participate (this type of randomized trials requires substantially larger sample sizes due the additional noise it introduces).

[8] Unless the trials are done in a large country such as China and India that allows randomization at the district or regional level. There are some examples of development trials in which this has been the case, although it is difficult to imagine similar trials in an OECD context (unless limited resources induce a sequential roll-out of a new intervention across regions, with the ordering of the roll-out being randomized).

optimization experiments can be used to improve national policies, ecosystem-level interventions and targeted support programs. They can both reduce the cost of delivering the policy and contribute to increase its impact.[9] An additional advantage is that results can often become available very quickly, making it possible to continuously iterate and inform immediate design choices. In the years we have been working in the field, we have not yet seen a policy intervention, even system-level ones, that would not benefit from embedding some randomized trials in the delivery.

For instance, while it is not possible to randomize the generosity of R&D tax credits, randomized trials can still be used to optimize how the scheme is delivered. Does providing personalized advice on how to apply increases take-up? Are there ways to reduce the number of ineligible claims? Does raising awareness about the scheme nudges companies to invest more in R&D? All these are testable questions. There are also similar questions that could potentially be tested about regulatory regimes that impact innovation.[10] Another example are infrastructure investments in an ecosystem, such as science parks or incubators. How to increase their use? How to maximize the benefits of co-location for tenants? How to run networks and/or seed new opportunities for collaboration between different actors? These are just a few of the questions that randomized trials can help to address.

There are also countless opportunities for optimization experiments in targeted support programs. A useful starting point is to map the user journey that participants follow throughout a program, spell out any questions and options in each of the stages, and then decide which of those are more likely to provide impactful insights and should be prioritized for testing. For instance, when the UK government's flagship business mentoring scheme was failing to recruit sufficient mentors, the UK business ministry (BEIS) used "nudging" trials experimenting with different language to increase recruitment rates, resulting in an additional 800 mentors recruited and achieving the policy target that otherwise would have been missed (to the surprise of the team involved, a quote by Adam Smith on the virtues of volunteering made the most difference).

Science and innovation funding processes are also a very fertile area for optimization experiments. Governments allocate billions through competitive funding calls, but there has been very little experimentation on how these processes are run. Are there ways to reduce the burden of the process? Can

[9] Impacts may not be measured directly but rather inferred from the assumptions embedded in the theory of change.
[10] For instance, whether providing more clarity about what is allowed and not under current rules supports innovation. A more ambitious example that falls in-between the different categories here would be testing the impact of the regulatory sandboxes that have proliferated across the world since the UK Financial Conduit Authority launched the first one in 2016. Not only it would generate evidence about the impact of this new policy instrument, but it would also provide some estimates of the costs that the regulatory regime may place on innovators.

we make the process more inclusive by encouraging more applications from women and minorities? Do review processes discourage novel or disruptive proposals? What biases impact decisions, and how can we prevent them? Many of these questions can be addressed with shadow experiments, setting up parallel shadow review processes without an impact on actual funding decisions,[11] while others are better tackled with full-fledged experiments.

Randomized trials can be totally theory-free, build on an existing theory, or try to test an actual theory and the hypotheses derived from it (mechanism experiments). While the latter trials can be more difficult to undertake, the results tend to be more rewarding, since they help to understand how innovation processes actually work. This knowledge is typically less context specific, and therefore has wider applicability to different settings. While an impact evaluation tells you whether something works or doesn't, a mechanism experiment can potentially answer why something works or doesn't, and prove or disprove assumptions about human and firms' behaviors. By shedding light on the underlying drivers of behaviors, mechanism experiments may provide insights that can be useful when designing different types of policies. Impact evaluations and mechanism experiments are not mutually exclusive, and the best randomized trials often try to combine both at once: measuring the impact of an intervention and understanding what causes the underlying behaviors.

Mechanism experiments can be conducted in the field (the "real world") or in a lab (typically in a university setting using undergrad students as subjects), or in "hybrid" settings, such as online platforms or using shadow processes. They can be framed around a particular innovation policy and seek to understand how and why individuals and firms react (or not) to it, or alternatively they can examine the management of innovation processes (whether in public or private institutions), which in turn can also help to inform innovation policy development.

Boudreau and Lakhani (2016) summarize in an earlier volume of this series the pioneering work that they have conducted at the Laboratory for Innovation Science at Harvard (and its predecessors the Crowd Innovation Lab and the NASA Tournament Lab). For instance, in several projects they have used randomized trials to understand how best to design innovation tournaments and contests, testing the impacts of competition and openness on innovators' effort, how its directed, and the resulting outcomes (as well as how these effects interact with the skills of participants and the complexity of the challenges being solved). More recently they have set up trials exploring other stages of the innovation process,

[11] A "shadow experiment" is conducted in a "real world" setting, but without impacting real decisions. For instance, it can be used to explore how decisions about which projects to fund would have changed if a different review process had been used in a competitive funding call, without actually changing the allocation of funding,

including the formation of scientific collaborations and peer review processes. Recent examples include Boudreau et al (2016), which finds that evaluators give lower scores to research proposals closer to their own areas of expertise and to highly novel research proposals, or Teplitskiy et al (2018), that provides evidence that female reviewers are more likely than male reviewers to be influenced by the views of other panel members.

4. What we are learning from randomized trials in innovation policy

It is still early to draw definitive conclusions from the policy experiments being undertaken in this area. Many are still on the field, some only have preliminary results, and others with findings have not yet been replicated in other contexts. But it is still useful to provide an overview of some of the emerging lessons from these trials, since they tackle some of the key policy questions that we face. This section does not aim to capture the full range of trials in this space (for a comprehensive overview see the online trials repository maintained by IGL). [12] Instead, it describes some examples that illustrate how randomized trials can be used to address policy relevant questions, taking a very broad definition of innovation policy. Many of these have received funding from the IGL Grants program and/or have been conducted by members of the IGL Research Network, which brings together over 85 researchers from around the world working in this space.

A large majority of the trials discussed here are evaluation experiments, seeking to answer whether a program works or not. Some are also mechanism experiments, testing the underlying drivers of behaviors, so a few examples are highlighted as well. This discussion does not cover optimization experiments, which focus on processes rather than outcomes and are not often codified.[13]

How do we get more and better ideas?

There is no innovation without new ideas or the creative re-combination of old ones. But how can we ensure that ideas easily flow and thrive? Are we as a society creating an environment that allows us to tap into all sources of new and interesting ideas? Unfortunately, whether we look at schools, universities, or businesses, the evidence suggests that we are missing out on many potential innovators and their ideas.

Increasing exposure to innovation

[12] The full database is available at www.innovationgrowthlab.org/igl-database.

[13] For an overview of the UK business ministry (BEIS) experience using optimization experiments, see this blog: www.innovationgrowthlab.org/blog/taking-first-steps-business-policy-experimentation.

In a highly influential paper Bell et al (2017) show that unless you are a top student from a high-income family, your chances of becoming an inventor or filing a patent are very low. They also find that growing up in an area with many inventors is a strong predictor of becoming one, and posit that lack of exposure to innovation when young is an important part of the story. They conclude that we are missing out on entire generations of inventors and their good ideas - the so-called "lost Einsteins". This is both detrimental to economic growth and a contributor to income inequality.

There are a number of programs that aim to tackle this challenge, but few of them have been rigorously evaluated (Gabriel et al, 2018). A new IGL trial led by the World Bank is testing out an online intervention to expose over 19,000 children in Latin America to STEM and entrepreneurship. In a previous IGL trial in Denmark, Moberg and Jørgensen (2018) find that a simple online entrepreneurship course for 9-graders could improve the sense of self-efficacy and their intention to pursue a career in entrepreneurship. These are only two of many approaches that could be trialed to cultivate innovative and entrepreneurial attitudes early on and address this important policy challenge.

Encouraging more people to participate

Encouraging individuals or groups who may not naturally consider themselves innovators or creative is another way to make innovation more inclusive and tap into new sources of ideas, but does it pay off? Two IGL trials suggest that it does. Both experiments were based around innovation contests, albeit in very different settings. The first trial was conducted with engineering and computer science students at a US university (Graff Zevin and Lyons, 2018), while the second took place at a large multinational in the Netherlands (Weitzel et al, 2019).

Despite the settings and research questions being slightly different, two of the findings were surprisingly similar. First, both studies showed it is possible to use messaging nudges and/or small financial incentives to encourage people to submit ideas into innovation contests. Secondly, and most importantly, there was no real difference in the average quality of the ideas submitted by someone who actively chose to participate in the innovation contest and someone who needed to be encouraged to join. In other words, encouraging more people to participate can lead to more ideas without decreasing their quality, and a small tweak in the process can be sufficient to make it happen. If we don't do it and instead rely only on self-volunteered contributors, we are missing out valuable ideas.

The trial in the Netherlands also looked at other ways to influence the quality of the ideas submitted, such as trying to widen the horizons of participants by showcasing successful projects from prior internal innovation contests. It turned out that in this setting this was counterproductive, making people less creative rather than more. Whether there are simple ways to make people more creative is something that another IGL trial in the UK is exploring, in this case looking at whether creativity can be trained through habit creation.

Another approach to encourage more people to take part in innovation and entrepreneurship is to rely on role models. Bechtold and Rosendahl Huber (2018) conducted an experiment which found that using female role models can be an effective way of fostering entrepreneurship among women. The power of female role models seems to persist even for actual entrepreneurs, as shown by an earlier trial in Chile, where it was found to be a cost-effective approach to boost income when compared to more expensive consulting services (Lafortune and Tessada, 2015).

Facilitating collaboration

Collaboration is an increasingly important component of any innovation process. The romantic idea of the sole inventor, with their "lightbulb moment", is today generally considered to be a myth. Instead, most scientists agree that complex challenges benefit from the combination of different expertise, knowledge, and backgrounds, and as a result teams have become more important (Wuchty et al 2007; Jones 2009).

However, we have little evidence on what the best ways to encourage collaborations are, both within and between universities and businesses. For instance, how important is physical proximity between researchers to facilitate collaboration?

Anecdotal evidence suggests that distance matters, and many new science labs have been built under the assumption that locating researchers from different fields under the same roof will unlock interdisciplinary research and open original research avenues. An IGL trial in Eastern Europe aims to test whether this is actually the case, by randomly distributing research groups within a large temporary research building, and tracking whether researchers physically located close to each other are more likely to collaborate.

A recent trial conducted at the Harvard Medical School suggests that close proximity, while important, is not enough. Specifically, an experiment by Boudreau et al (2017) finds that there are substantial search costs that affect matching between scientific collaborators, even when these are located in the same institution. The experiment also demonstrates how a simple low-cost intervention creates new collaborations that otherwise would not exist. Specifically, bringing together scientists working under the same roof to talk about their ideas with each other in a 90-minutes structured information sharing session increases the probability of grant co-application of a given pair of researchers by 75%. Both these trials are rare examples of applying the scientific method to science policy.

Collaborations between researchers and businesses are also a source of new ideas, but this too is an area where there is overwhelming agreement that we are missing out on many opportunities. A number of IGL trials are already looking at different ways of addressing this challenge and we are also planning further work. One of the instruments that has become popular in recent years is innovation vouchers. Their aim is to nudge SMEs to engage with universities and other knowledge providers by providing small vouchers (typically around $5,000 to $15,000) to buy research services from them. A trial by the Dutch government found that innovation vouchers were effective at creating new collaborations between SMEs and universities (Cornet et al, 2016). While this trial had not been originally conceived as a mechanism experiment, [14] it still produced useful insights about the underlying mechanisms. Specifically, the trial found that these new collaborations did not continue once the subsidy stopped, suggesting that the main barrier to SME-university collaboration was not lack of information or connections (the underlying assumption behind the policy), but rather a much more fundamental one. An ongoing trial with the UK's national innovation agency (Innovate UK) is expected to shed more light on this question.

How can we support entrepreneurs and business to scale and adopt new ideas?

Good ideas are not of much use unless they are put into practice, scaled up and widely adopted. This is why there is a long list of programs and policies to support this process, ranging from entrepreneurship training initiatives, accelerators and other startup support programs at one end, to innovation grants, SME finance schemes, business support or tech adoption programs at the other. Despite the substantial budgets involved,[15] the impact most of these programs have is unclear, and we don't know either whether changing their design would make them more or less effective. A growing number of trials are trying to provide some answers.

Training entrepreneurs and supporting startups

Entrepreneurship has become increasingly popular, and this is reflected in the growing number of universities, private providers and governments offering entrepreneurship training today. What remains

[14] In fact, this trial had not been originally conceived as a randomized trial. However, excess demand for the first round of the program led the government to use a lottery to allocate the vouchers, and the same system was used in subsequent rounds.

[15] For instance, European governments spend every year ca. $170 billion in public programs to support entrepreneurs and businesses to innovate and grow (Firpo and Beevers, 2016).

unclear is which type of training best fulfills the needs of entrepreneurs, and what those needs actually are.

A recent trial by the World Bank in West Africa tries to address both of these questions (Campos et al, 2017). It shows that a personal initiative training approach for microentrepreneurs, which teaches a proactive mindset and focuses on entrepreneurial behaviors,[16] can be much more effective than teaching them formal business skills, such as marketing or financial management. Specifically, the psychology-based training increased firm profits by 30% and payed for itself within one year (compared to a non-statistically significant 11% increase in profits for traditional business training). The effect was even stronger for female-owned businesses. In addition to demonstrating the impact of this particular training program, this trial also sheds some light on the much larger question on whether an entrepreneur is "born or made", by demonstrating that some entrepreneurial attributes are not totally innate, can be taught and make a difference in entrepreneurial performance.

A number of IGL trials are also considering similar questions. A forthcoming trial in Jamaica is comparing traditional business skills classes with classes on personal initiative and persistence (Ubfal et al, 2019). Another, in Italy, is teaching entrepreneurs to become more experimental by teaching them to use hypothesis-based experiments to assess the viability of their business idea(s) and evaluate the effect of their strategies. The results from their pilot study show that the training had a positive effect on startup performance (Camuffo et al, 2019), and the intervention is now being tested with larger samples in Italy and the UK. As some of the prior examples, these trials not only test the impact of a particular training module, but by doing so also provide supporting evidence for entrepreneurship theories that see entrepreneurship as a structured discovery process (such as lean startup methods).

This structure can be self-imposed by the entrepreneurs themselves (as above), or alternatively imposed on them by others. Independence and not having a boss are often cited as reasons why entrepreneurs decide to start their own business, but preliminary findings from another IGL trial suggest this can work against them (Leatherbee, 2019). In the trial, which took place in a large accelerator program in Latin America, all entrepreneurs participated in monthly meetings, but those in the treatment group were required to reflect on the success of the tasks they had committed to at the previous meeting and share the tasks they planned to execute before the next meeting. Early results suggest that the introduction of these additional accountability structures helped to improve startup performance.

[16] Including self-starting behavior, innovation, identifying and exploiting new opportunities, goal-setting, planning and feedback cycles, and overcoming obstacles.

Accountability is important, but even simple feedback without strings attached can make a difference. Government agencies are often reluctant to share detailed feedback on the proposals that they review, for the fear of opening the door to lots of complaints. The question is whether we are losing out from that. A trial showed that giving startups in the Startup Chile program the feedback collected as part of the selection process increased both external fundraising and survival probability (Wagner, 2017). One of our IGL partners is now trying to replicate this trial with one of its programs, in order to decide whether it is worth sharing the detailed feedback that they are collecting in the process of reviewing funding proposals.

Supporting SMEs innovation and productivity

Reversing the productivity slowdown requires getting more SMEs to innovate and/or adopt new technologies and production methods (Andrews et al, 2016), but the best way to achieve is still an open question.

There is a broad spectrum of targeted interventions that have the potential to increase firm productivity, some very intensive and others much more light-touch. Recent trials demonstrate that they both can work. Bloom et al (2013) conducted a trial involving 17 poorly managed Indian textile firms. All of them were given customized recommendations for improving management practices, but only the "treatment" plans received additional "high-grade" management consultant support during several months to help them implement the recommendations. The consultancy made a substantial difference in the uptake of the recommended management practices and led to significantly larger performance improvements. The intervention was not cheap, but the productivity gains more than offset the cost. The authors also followed up several years later and found that many of the effects persisted (Bloom et al, 2018). Another recent example, trialing less intensive consulting services over a year for a larger sample of 432 SMEs in Mexico, also found strong effects on employment, productivity and return on assets (Bruhn et al, 2018). In addition to evaluating the impact of their respective interventions, both of these trials demonstrate the importance of managerial capital and contribute to a much larger question: Should governments provide public-funded support to profit-maximizing businesses to encourage them to adopt practices and technologies that increase their own profitability? Is providing information about them enough (i.e., making the "unknown unknowns" known), or do businesses also need additional hand-holding in the form of intensive support to do what supposedly is good for them? If so, why and when? Both of these trials suggest that information is not enough, even if they don't fully answer the "why" question. The series of trials on management and technology adoption being funded the UK government's Business Basics program discussed earlier will hopefully shed some additional light on this important question.

There is also evidence at the other end of the spectrum, demonstrating that light-touch interventions can work. A recent trial in China shows that a simple low-cost intervention that got small businesses to meet in small groups once a month for a year led to increased sales and profits (Cai and Szeidl, 2018). The impact was much larger than for much costlier interventions (firm sales went up by 8%), so the program was scaled up and a similar program is set to be trialed in the UK. As part of the study, the researchers also tested the concrete channels behind the impact, and found that the meetings facilitated peer-learning between firms and improved supplier-client matching (the effect was larger for firms in groups with better peers).

Instead of in-kind support, other programs directly provide funding to SMEs, either small or large amounts. For instance, Nesta's Creative Credits trial successfully used small vouchers to encourage SMEs to work more closely with creative suppliers, although these new relationships didn't last in the long term (Bakhshi et al, 2013).[17] As discussed above, many other vouchers schemes are based on a similar logic, and the jury is still out on what their ultimate impact is.

While small vouchers schemes are popular, much more budget is allocated to funding large R&D and innovation grants. Do these types of large grants replace existing investment that firms would have made in any case, or do they mostly lead to new activity? An IGL trial led by the World Bank in Latin America is trying to answer this question. Randomizing innovation grants as large as $250,000 would not be a very popular policy decision, so all the funding applications that are scored highly by all reviewers will get the grant, while those that everyone scores poorly will not. Funding for applications in which there are disagreements between the different reviewers will be randomized, and the impact tracked. The implicit, yet untested, assumption is that value for money is higher for the applications with the top scores, although it could well be that these are precisely the ones that companies or investors would have funded in any case.

This trial in Latin America is also trying to understand who is better at making decisions about which companies to support. As discussed earlier, this fits into a much wider question, namely how do we run selection processes to allocate public research and innovation funding. This is an area that is ripe for experimentation in which at IGL we are planning additional work with some of our partners.

[17] This study had been originally conceived as a mechanism experiment that sought to test how important creative inputs are for innovation, with the voucher intervention being developed as the instrument to test this hypothesis (although the resulting evaluation led to this voucher program being replicated in other locations).

5. Overcoming the barriers to policy experimentation

The barriers to experimentation

Despite the many benefits of experimentation, policy organizations often find it difficult to take it on (Breckon, 2015). A range of barriers, both real and perceived, slowdown the adoption of randomized trials in innovation policymaking. Many are common with other types of evidence, and relate to the challenges of evidence-based policy making. Others are specific to the use of randomized trials.[18]

In order to understand them and help inform IGL's future work, in 2016 we conducted a small survey of policy makers in this space.[19] We inquired about the main barriers to evidence use and randomized trials specifically.

Some of the most common barriers to evidence use are political, relating to the policy cycles or competing political priorities. For instance, the pressure to make policy decisions before rigorous evidence emerges. Limited availability of rigorous evidence was also highly mentioned, as was insufficient demand for evidence. Policymakers often believe their opinion is correct and do not feel they need better evidence on their programs' impact.

When asking about randomized trials, most of the potential barriers proposed in the survey were considered to be very important or important by respondents, highlighting the multi-faced nature of the challenges that need to be overcome to increase policy experimentation.

Barriers fall into different categories. Concerns with public reactions to randomization and fear of negative results are very frequent. Others are related to lack of knowledge, such as limited awareness of the value of randomized trials (particularly among senior officials) or insufficient skills to conduct trials. A related barrier are budgetary constraints and missing organizational processes and structures. Finally, the perception that randomized trials are not feasible or timely also limits their adoption (although as discussed in the prior section these perceptions do not necessarily match reality).

It is useful to address three common misconceptions about randomized trials that are often mentioned by policymakers as reasons not to adopt them. The first one is that randomized trials are "too expensive". Although large clinical trials are notorious for their costs, not all trials need be overly pricey. Often the

[18] A third group would include many of the criticisms that are made to the use of randomized trials that equally apply to many other evaluation methods, but which are only raised when someone proposes running a randomized trial.
[19] Results available at www.innovationgrowthlab.org/blog/barriers-experimentation-survey-results.

most expensive part in a trial is the program itself - a cost which the organization is presumably incurring in any event. Depending on which data is needed, data collection can also take up many resources, but this is true of any type of evaluation regardless of the method used. Increasingly there are alternative data sources that reduce the need to undertake expensive surveys, substantially reducing data costs (even if surveys can still be necessary depending on the outcomes of interest). Administrative data sources are becoming more available, with governments increasingly more willing to open or share their administrative data, alongside increasing investments in data cleaning and matching. Moreover, the emergence of "big data" allows tracking some potential outcomes from our digital footprints, for instance using web scraping.

Beyond program and data collection costs, the actual costs of running a randomized trial are relatively low. Compared to more traditional approaches, randomized trials have a higher upfront cost in terms of design and analysis, but there are some advantages as a result. Setting up a trial forces organizations to carefully look at the data from the beginning, invest time to understand the actual problem the program aims to address, develop a logic model or theory of change that really breaks down the channels of impact and the underlying program assumptions, and put in place monitoring systems from the outset. The structure trials impose its beneficial even on its own, since it leads to better designed programs and more careful execution. In addition, many academic researchers are willing to collaborate with policymakers at little or no cost on the design and analysis of randomized trials, since a well-executed trial tackling an interesting research question in this field can easily get published in a top academic journal.[20] Ultimately, there are many missed opportunities in the field of innovation policy to run relatively cheap trials whose findings would pay for themselves (either by saving their cost if the program proves to be ineffective, or by significantly improving their effectiveness with simple tweaks).

The second misconception is that "it is unethical to withhold support to some participants". This is a criticism that is more frequent in organizations not engaged in experimentation than in those undertaking trials, and it often mixes ethics and the fear of backlash from the public.[21] It is definitely true that careful attention should be paid to the ethical implications of a randomized trial, and the specific context matters. An implicit assumption behind this criticism is that trials involve denying some potential recipients an

[20] Connecting policymakers looking for support with researchers interested in collaborating on randomized trials is one of the activities of IGL.

[21] With regards to the latter, in order to reduce the risk of backlash, it is important to develop a careful communications strategy that demonstrates the value of experimentation and gets buy-in from the key stakeholders involved.

intervention that would benefit rather than harm them. However, this cannot be taken for granted.[22] Rolling out programs without knowing whether they are beneficial or harmful is a risk worth preventing. This is why trials are widely accepted in much more difficult contexts, like testing new life-saving drugs.

Even for those interventions for which "harm" is extremely unlikely, there is still an "ethical" case to be made in favor of experimentation (rather than against it). Spending taxpayers' money on a program that is ineffective deprives more effective programs of funding, so trials can help elucidate whether we are making a good use of limited public resources. In many circumstances, moreover, a randomized trial does not require that the control group receives nothing at all - often different versions of the same program are pitted against each other, with all participants receiving some of the intervention in some form. Alternatively, when programs are rolled-out progressively (rather than for everyone at once) due to budget or capacity constraints, the order in which the program is introduced can also be randomized. In this case, no one is denied the program, and those required to wait may ultimate benefit by getting a more developed and therefore more effective intervention.

A related misconception is that "it is unfair to use a lottery to select participants". There is a fear that undeserving applicants may benefit from a program, while those that would benefit the most do not. Budgets are often insufficient to support all deserving applicants, so the question is what is the most appropriate method of allocating limited funding. In some circumstances a lottery can be a fairer and more efficient approach to select participants than other systems frequently used, such as first-come first-serve criteria or proposal scoring systems.

An implicit but often untested assumption in selection processes is that projects that score best are those that should be selected. However, in some contexts highly-scoring proposals may be those for which additionality is lowest, since they may get funded anyway by the private sector. In addition, review scores may be noisy and reward the best proposal writing consultant rather than the most promising project. For instance, the evidence suggests that traditional proposal scoring systems, such as those used in science funding programs, are effective at filtering out poor proposals, but they fail to identify the best ones (Graves, 2011; Boudreau et al, 2016). While politically it might be more difficult to justify a lottery for a costly intervention than a low-cost one, the rationale for randomization does not depend on the magnitude of the intervention (as long as the sample size is large enough). Ultimately, lotteries can be designed in different ways to accommodate an organization's criteria. For instance, randomized trials do not require

[22] For instance, a trial examining an entrepreneurship support program in the US found out that the quality of the training was so weak that, rather than helping firms, if anything the impact on business performance was the opposite, although results were not statistically significant (Fairlie et al, 2015).

providing funding to undeserving applicants, since those can be screened out prior to conducting the lottery. Similarly, best ranked applications might be funded directly, with the lottery being used instead to select among similarly-ranked middle-tier applications.

How IGL has been working to increase policy experimentation

The barriers to policy experimentation are deeply ingrained in the day-to-day of innovation policymaking, so changing the status quo is not done overnight. Embedding a culture of experimentation across economic ministries and innovation agencies around the world will take time. It requires raising awareness of the value and feasibility of policy experimentation. Identifying early champions within governments. Helping them set up their first trials, often small ones, which in turn make it easier for them to build internal coalitions to undertake larger and more impactful trials. Getting the resulting evidence used and successful programs scaled-up. And, lastly, sustaining this change until it becomes part of the norm, institutionalized in processes, instruments and budgets.

Through our work we have tried to simultaneously tackle the different barriers hindering policy experimentation: increasing awareness, developing skills, providing funding,[23] advising governments, disseminating knowledge, creating open resources, connecting networks and facilitating peer learning.

While it will be a long process until the full impact starts to materialize and experimentation becomes "normalized", we are already starting to see some progress. When IGL was launched, few innovation policymakers across OECD countries had seriously considered setting up policy experiments in this space. When confronted with this idea their response was often quite dismissive: "It can't (or shouldn't) be done".

Since then, we have been fortunate to count as partners some of the leading innovation agencies and ministries in the world.[24] As a result, today over 15 government agencies in 10 OECD countries have launched or are actively considering policy experiments in this space, with several developing in-house capacity to undertake trials. We've supported over 55 trials, have IGL partners or projects in 26 countries, and have worked with more than 25 organizations to help them become more experimental. We have also started to build a global community bringing together policymakers and researchers who share this

[23] Through the IGL Grants program, thanks to the generous funding from Nesta, the Kauffman Foundation and the Argidius Foundation.

[24] The IGL partnership has included the following innovation agencies and government ministries: ACCIÓ – Catalan Agency for Business Competitiveness, the Australian Department of Industry, the Danish Business Authority, Design Singapore, Austria's FFG, Innovate UK, Innovation Norway, the Ministry of Economic Affairs of the Netherlands, Scottish Enterprise, the Swedish Agency for Growth Analysis, Finland's TEKES, and the UK Department for Business, Innovation and Skills.

mission, and our events, capacity-building workshops and online resources have reached thousands of policymakers from close to 50 countries.

We are also starting to see the first steps towards institutionalizing policy experimentation through experimentation funds. New ideas for support programs are everywhere in the ecosystem, not just in government buildings. So an important question is what mechanisms governments have to first identify them, and then distinguish between program ideas that should be scaled up vs. well-intentioned but ineffective efforts. Experimentation funds can be a solution. They provide funding to test innovative support schemes in exchange for rigorous evaluation. In other words, they are a mechanism to identify, test and support the most promising ideas for support programs, often coming from organizations that are much more closely engaged with businesses, and not just from the usual suspects.

Experimentation funds have been set up in the past in policy areas such as education (through the UK's Education Endowment Foundation) and youth programs (as with France's Fonds d'Experimentation pour la Jeunesse). Building on their experience, at IGL we developed a blueprint on why and how to set them up, and are collaborating with governments in designing and delivering them.

Both the European Commission and the UK government have recently taken up this idea. The Commission has launched the first funding call for policy experiments targeted at innovation agencies across Europe. In the UK, as part of the new Industrial Strategy, the government has set up the Business Basics program. Being delivered in partnership with IGL, Business Basics will fund a range of experimental projects that test innovative ways of encouraging SMEs to adopt new technologies and management practices. For instance, one of the pilots is testing how to get SMEs to adopt AI-based technologies.

Some lessons about the experimentation journey

Through our engagement we have learnt some lessons about how best to support policymakers to embark into the experimentation journey. When trying to raise awareness about experimentation, simple examples and stories can be very powerful. The lack of relevant examples can make the idea of experimenting somewhat abstract. Showcasing experiences from other governments that closely relate to someone's job and challenges not only helps to inspire them, but also confers some "protection" or "cover" and can give them the confidence to propose a different path to reluctant senior managers or ministers.

How experimentation is framed also matters. Policymakers are much more receptive to the idea of using randomized trials to test ways to improve the effectiveness of their programs, than being told that you are there to independently evaluate their program to find out whether it works (unfortunately a much more threatening proposition).

Buy-in from junior and senior policymakers is a necessary condition, which requires aligning what you would ideally like to test with what they actually care about. This means being opportunistic and building on the ideas and problems that policymakers and program managers already have. Compromises are often needed, so it helps being mindful of what can and cannot be achieved, and trying to understand and address their concerns (rather being dogmatic). Face-to-face interactions with the organization's staff through meetings and dedicated workshops tend to work best.

Policy experimentation goes beyond randomized trials, so policymakers can achieve the best results when they think experimentally throughout the policy cycle, using a range of methods to explore new and innovative solutions to policy challenges rather than only focusing on randomized trials. In other words, avoid being a hammer in search of nail, and use randomized trials at the right stage, ideally once the intervention has been prototyped (although this can be difficult in compressed policy cycles).

Getting an organization to the point of running their first randomized trial is not easy. It is worth using an incremental approach, identifying the path of least resistance and developing a portfolio of opportunities. In our experience, a useful starting point is to run messaging trials - behavioral experiments to find out what language is most suitable to achieve a certain goal, such as convincing firms to take up a program (this may be as simple as using A/B testing to experiment with different newsletter formats and the specific language used on them). As these are relatively close to business as usual, they can typically be run under-the-radar without requiring senior approval, and can contribute to building an internal coalition for experimentation involving several teams from the organization. If the results are counterintuitive and surprising, even better.

A portfolio approach is important because most trial opportunities will fail to reach the field. The larger and more ambitious a trial is, the more likely it is that it will be stopped by someone at some point during the development and approval processes. So hedging your bets is useful, rather than focusing only on a full-fledged impact evaluation of a new program (particularly a very popular one). One way to do this is to work on a number of small-scale pragmatic trial options focused on testing improvements to existing programs before rolling them out to all participants. This could be changes in the delivery mode (e.g., online vs. face-to-face), adjustments to the internal processes, or incorporating an add-on program on top of an existing scheme. Unless you are in a very flexible and open-minded organization, large-scale trials that substantially deviate from business as usual only become feasible once there is sufficient buy-in at different levels and an internal coalition to support it.

Governments can engage in experimentation by running trials in-house, but also by supporting trials done elsewhere that are also aligned with their mission. Experimentation funds, discussed above, are one way of doing that. But there are others, from explicitly signaling in existing funding programs that randomized trials are welcome or eligible, to setting up small calls to fund specific trials. This can help organizations not only to access new ideas and contribute to build the evidence base, but also to observe up close how trials are conducted and in the process progressively get ready to do them also in-house.

Once an opportunity to set up a randomized trial finally arises, it is important to make sure that the trial is designed and executed well, to get robust answers, but also to demonstrate to the wider organization the value that they can derive from trials. At IGL we have collected many of the lessons about how to conduct randomized trials in this field on an experimentation guide and an online toolkit (Edovald and Firpo, 2016; IGL, 2017)[25], and regularly provide advice and support to policymakers to help them in this process, something recommended particularly if an organization is new to the world of experimentation.

6. Concluding remarks

While embracing policy experimentation is a substantial change from business as usual, starting the experimentation journey requires only a few small steps. Many fear that experiments are too complex and disruptive of the status quo - assuming any trial must set out to randomize large sums of funding or radically alter the way a program is run. Yet as this paper has shown, there are many ways to become experimental, and also many reasons to do so. Improving the evidence base is one of them, but an often overlooked benefit of experimentation is how it encourages organizations to become more agile and innovative, continuously searching for new ideas to test rather than defaulting to the status quo.

Given the range of experimentation approaches in existence, it is important to select the appropriate method, which depends, among others, on the question being asked, what we know about potential solutions, their stage of development, the level at which the intervention will be implemented or the time it will take to show results. More often than not, these methods can be used as complements, not

[25] Both available here: www.innovationgrowthlab.org. See also the World Bank's Development Impact Blog Series (blogs.worldbank.org/impactevaluations/) for a large number of blogs tackling many of the practical issues involved in running randomized trials with firms.

substitutes - for instance using design approaches to better develop the blueprint of a program, and a randomized trial to test its effects.

Experimentation lowers overall policy costs, because, despite investing a little more upfront in learning and evaluation, experiments allow policymakers to “weed out” ineffective programs early on, potentially saving taxpayers from footing the bill. It can also help increase the impact of existing programs - by constantly testing tweaks in the way they are delivered. Experimenting with new programs can strengthen their design from the outset, by testing different versions or components of a program, and understanding how they fit together. When it comes to deciding which programs to scale, randomized trials are especially well suited to inform decisions, because their results typically come in the form of a robust quantitative estimate that can be easily used to do a cost-benefit analysis.

An often overlooked benefit of experimentation is how it can help de-risk the process of exploring new ideas and challenges. By starting small and testing effectiveness early, experiments can in fact make it easier for risk-averse organizations to sample novel approaches and venture into more innovative fields, without having to commit large amounts of resources (and thus reputation) in the process. As with any other innovation, some of these will undoubtedly fail, but these are “good failures” that create useful knowledge and prevent unnecessary “bad failures” from happening. In other words, they are small-scale, controlled, and ultimately unavoidable if we want to learn about what works in an uncertain and complex world.

Experimentation is only one of the ingredients for delivering good innovation policy. Better use of data could also greatly contribute to develop more effective policies, if governments were more willing to open their in-house administrative data to the wider researcher community (really the low-hanging fruit to improve the evidence base), and sought to fully exploit the vast opportunities of “big data”. Lastly, good judgement will always be required, since in an uncertain world where information is incomplete, the evidence base can only take us so far. But it could take us much further if governments made it a priority.

Overcoming the challenges we face will require not only new ideas, but also learning whether they work. At IGL we believe that becoming experimental can help policymakers ask the right questions and get better answers.

7. References

Andrews, D, C Criscuolo and P Gal (2016). *The Best versus the Rest: The Global Productivity Slowdown, Divergence across Firms and the Role of Public Policy*. OECD Productivity Working Papers, No. 5, OECD Publishing, Paris.

Ansell, C K, and M Bartenberger (2016). *Varieties of experimentalism*. Ecological Economics, Volume 130, October 2016, Pages 64-73.

Azoulay, P (2012). *Turn the Scientific Method on Ourselves*. Nature. Volume 484, pages 31–32.

Bakhshi, H, J Edwards, S Roper, J Scully, D Shaw, L Morley, and N Rathbone (2013). *Creative Credits: A Randomized Controlled Industrial Policy Experiment,* Nesta.

Bakhshi, H, A Freeman, and J Potts (2011). *State of Uncertainty: Innovation Policy through Experimentation*. Nesta Provocation no 14.

Bechtold, L A, and L Rosendahl Huber (2018). *Yes I Can! - A Field Experiment on Female Role Model Effects on Entrepreneurship,* Academy of Management Proceedings, Vol. 2018:1.

Bell, A, R Chetty, X Jaravel, N Petkova, and J Van Reenen (2017). *Who Becomes an Inventor in America? The Importance of Exposure to Innovation,* NBER Working Paper No. 24062.

Bloom, N, B Eifert, A Mahajan, D McKenzie, and J Roberts (2013). *Does Management Matter? Evidence from India*, The Quarterly Journal of Economics, Vol.128:1, pp. 1–51.

Bloom, N., A. Mahajan, D. McKenzie, J. Roberts (2018). *Do management interventions last? Evidence from India*, NBER Working Paper 2424*9*

Boudreau, K, T Brady, I Ganguli, P Gaule, E Guinan, A Hollenberg, and K Lakhani (2017). *A Field Experiment on Search Costs and the Formation of Scientific Collaborations*, Review of Economics and Statistics, Vol. 99:4, pp.565-576.

Boudreau, K J, E Guinan, K R Lakhani, and C Riedl (2016). *Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance and Resource Allocation in Science*. Management Science 62, no. 10 (October 2016).

Boudreau, K J, and K R Lakhani (2016). *Innovation Experiments: Researching Technical Advance, Knowledge Production and the Design of Supporting Institutions.* In *Innovation Policy and the Economy, Volume 16*, edited by W R Kerr, J Lerner, and S Stern, 135–167. National Bureau of Economic Research, and University of Chicago Press.

Breckon, J (2015). *Better Public Services Through Experimental Government*, Nesta.

Bruhn, M, D Karlan, and A Schoar (2018). *The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico*, Journal of Political Economy Vol.126:2, pp. 635-687.

Cai, J, and A Szeidl (2018). *Interfirm Relationships and Business Performance*, The Quarterly Journal of Economics, Vol.133:3, pp. 1229–1282.

Campos, F M L, M Frese, M, Goldstein, L, Iacovone, H C Johnson, D J Mckenzie, and M Mensmann (2017). *Teaching personal initiative beats traditional training in boosting small business in West Africa*, World Bank Group.

Camuffo, A, A Cordova, A and Gambardella (2019). *A Scientific Approach to Entrepreneurial Decision-Making: Evidence from a Randomized Control Trial*, Management Science, Forthcoming.

Cockburn, I, R Henderson and S Stern (2018). *The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis*. In *The Economics of Artificial Intelligence: An Agenda*, edited by A K Agrawal, J Gans, and A Goldfarb. National Bureau of Economic Research, and University of Chicago Press.

Cornet, M, B Vroomen, and M van der Steeg (2006). *Do Innovation Vouchers Help SMEs to Cross the Bridge Towards Science?*. CPB Discussion Paper, no. 58.

Criscuolo, C, R Martin, H G Overman and J Van Reenen (2019). *Some Causal Effects of an Industrial Policy*, American Economic Review, vol 109(1), pages 48-85.

Deaton, A and N Cartwright (2018), *Understanding and misunderstanding randomized controlled trials*. Social Science & Medicine, Volume 210, August 2018, Pages 2-21.

Edler, J, P Shapira, P Cunningham, and A Gok (2016). *Conclusions: Evidence on the Effectiveness of Innovation Policy Intervention*, in J Edler, P Cunningham, A Gök and P Shapira (eds), Handbook of Innovation Policy Impact. Eu-SPRI Forum on Science, Technology and Innovation Policy series, Edward Elgar Publishing, Cheltenham, pp. 543-564.

Edovald, T and T Firpo (2016). *Running Randomised Controlled Trials In Innovation, Entrepreneurship And Growth: An Introductory Guide*. Innovation Growth Lab. London.

Fairlie, R W, D Karlan, and J Zinman (2015). *Behind the GATE Experiment: Evidence on Effects of and Rationales for Subsidized Entrepreneurship Training*. American Economic Journal: Economic Policy, vol. 7(2), pages 125-61.

Firpo, T, and T Beevers (2016). *As Much as €152 Billion Is Spent Across Europe Supporting Businesses: But Does it Work?*, IGL. Mimeo.

Gabriel, M, J Ollard, and N Wilkinson (2018). *Opportunity Lost: How inventive potential is squandered and what to do about it*, Nesta Report.

Graff Zevin, J, and E Lyons (2018). *Can Innovators Be Created? Experimental Evidence from an Innovation Contest*, NBER Working Paper No. 24339.

Graves, N (2011). *Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel*, BMJ 2011(343).

IGL (2017). The IGL Experimentation Toolkit. Innovation Growth Lab. London. Available at: http://toolkit.innovationgrowthlab.org/home

Jones, B F (2009) *The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?* The Review of Economic Studies, Volume 76, Issue 1, 1 January 2009, Pages 283–317.

Lafortune, J, and J Tessada (2015). *Improving financial literacy and participation of female entrepreneurship in Chile*, Final Report to Global Development Network CAF/GDN Project.

Leatherbee (2019). *Better Flee from Freedom? Testing the Effects of Structured Accountability on New Venture Performance*. Mimeo,

Moberg, S, and C Jørgensen (2017). *Entrepreneurial Role Models and Online-based Entrepreneurship Education: Results from an Ongoing RCT*. Danish Entrepreneurship Foundation. Mimeo.

Svensson, P, and R K Hartmann (2017). *Policies to Promote User Innovation: The Case of Makerspaces and Clinician Innovation*, MIT Sloan Research Paper No. 5151-15.

Teplitskiy, M, H Ranu, G Gray, E Guinan, and K Lakhani (2018). "Gender, Status, and Willingness to be Wrong: A Field Experiment in Scientific Peer Review." LISH, Mimeo.

Ubfal, D, I Arraiz, D Beuermann, M Frese, A Maffioli, and D Verch (2019). *The Impact of Soft-Skills Training for Entrepreneurs in Jamaica*, Mimeo.

Wagner, R (2017). *How Does Feedback Impact New Ventures? Fundraising in a Randomized Field Experiment*. Mimeo

Weitzel, U, C Rigtering, and A Fenneman (2019). *Increasing quantity without compromising quality: How managerial framing affects intrapreneurship*, Journal of Business Venturing, Vol 34 Issue 2 (March 2019) pp 224-241.

Wuchty, S, B F Jones, and U Brian (2007). *The Increasing Dominance of Teams in Production of Knowledge*, Science 316, p. 1036