Chapter 8.

Policy experimentation in complex education systems

Patrick Blanchenay, *University of Toronto*

Tracey Burns, *Centre for Educational Research and Innovation, OECD*

Complexity is increasing in education – in governance arrangements, in the numbers of stakeholders and in the availability and use of evaluation and other accountability data. These changes call for moving away from a traditional policy cycle towards one which can evolve and adapt with our systems in order to govern them effectively. One tool of this new kind of governance is policy experimentation.

This chapter suggests avenues to make experimentation a more effective instrument for policy making in a complex environment, and demonstrates that a tension exists between properly evaluating the effects of narrowly-focused experiments and translating these results into the broader network in which every stakeholder is embedded. It suggests that a good balance can be struck by experimenting at a suitable scale, and moving towards what is called ecosystem experimentation.

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Introduction

The increased devolution of responsibilities and tasks from central governments to lower levels of government and local authorities, combined with a trend towards increasing accountability, has contributed to make education systems in the OECD countries more intricate and more complex. This trend has happened both vertically, with the interaction of different levels of governance, as well as horizontally, with increased involvement of a wider range of actors in all processes.

This increased complexity – in governance arrangements, in the numbers of stakeholders, and in the availability and use of evaluation and accountability data – calls for a new approach to governance. Education systems are in fact *complex systems* – that is, networks of interdependently linked actors whose actions affect all other actors, and which evolve, adapt, and reorganise themselves. Complex systems do not work in a linear manner but rather exhibit a series of well-defined characteristics: tipping points, feedback loops, path dependence and sensibility to local contexts (Bryne, 1998).[1]

Understanding complexity is an important point for policy making and governance in general, as complex systems cannot be successfully governed or steered with simple, linear mechanisms. In complex environments, strategies must be developed that take into account the dynamics and interdependency of the system. Simply devolving power to local authorities will not improve the functioning of the system unless it is also accompanied by attention to the connections and interactivity present, and space is made to facilitate and use the constant feedback that is required to guide complex systems. However, in complex environments such as education in which a multitude of actors are collaborating through formal and informal channels, the sheer amount of feedback and interactivity can seem impossible to navigate effectively.

Policy experimentation has been suggested as one strategy for dealing with such complexity. Policy experimentation can be defined as "a purposeful and coordinated activity geared to producing novel policy options that are injected into official policymaking and then replicated on a larger scale" (Heilmann, 2008b). In practice this implies the deliberate implementation of a new programme or practice on a small scale, targeting a selected number of schools or districts, with the intention of evaluating the effectiveness and possible scaling up to a wider level if effectiveness is demonstrated.

The present chapter provides a rationale for the use of policy experimentation in the context of complex education systems, and shows under what circumstances it could be useful. First, the chapter outlines broad governance challenges posed by complexity. It then defines more precisely what is meant by experimentation, what we can learn from it in a complex environment, and places experimentation as a useful tool in the debate between big package interventions and focused incremental reforms. In a third section, the chapter shows how experimentation can be tailored to account for complexity by choosing a suitable scale that we call "ecosystem experimentation", and by adopting the policy cycle to account for such complexity. It then looks at the risks involved with experimentation, and highlights the necessity of building educational systems capable of taking risks and can learn from both success and (importantly) failure.

Finally, the chapter suggests avenues to make experimentation a more effective tool for policy making in a complex environment. A tension exists between properly evaluating the effects of narrowly focused experiments and translating these results into the broader network in which every stakeholder is embedded. It suggests that a good

balance can be achieved by experimenting at a suitable scale, and moving towards what is called *ecosystem experimentation*.

Governance challenges in a complex environment

Across the OECD, education governance has in general moved away from hierarchical governance systems towards more complex environments in which a multitude of actors collaborate through formal and informal channels. In such contexts, the successes and failures of students and schools depend on a multitude of interdependent actors, who all play a part in moving the system forward. This interdependence poses several challenges for the governance of education systems that are critical for this discussion:

1. The traditional linear approach to policy-making may not be best suited to operate in such an environment.
2. Successful governance requires the co-operation of actors with different motivations, outlooks and time horizons.
3. Information is more plentiful and also more scattered, both in its production and its consumption.

The traditional policy cycle is inadequate

The complexity of educational systems means that the traditional policy cycle is not able to capture the interplay and dynamics characteristic of modern arrangements. This is not radical or startling news: in fact, there have been numerous critics describing the inadequacy of the traditional policy cycle, and not just in education. Clay and Schaffer (1984) made this argument in relation to agricultural policy 30 years ago, and these comments have been broadened and deepened in almost all public sectors since. Hallsworth et al. (2011: 38-44) provide a detailed breakdown of why this is so:

- "Policy making does not occur in distinct stages". Instead, problems and potential solutions often emerge together, rather than sequentially.
- "Policies need to be designed" properly to reach their goals, but it is hard to design something perfectly without a trial or implementation attempt to refine it. This problem is magnified by the fact that feedback loops in a complex system may amplify the unpredictability present and the whole set-up of links in the system shifts as the experiment progresses, thereby creating potentially unintended consequences.
- "Policy making is often determined by events". Politics and media attention are important forces affecting potentially powerful stakeholders; this may sometimes lead to decisions taken for political reasons, often with an emphasis on short-term results. In the age of social media, real time examples of the power of different stakeholders (for example parents or students) in changing the terms of a debate or calling into question the adequacy of a policy response to a particular problem can be powerful forces in politics and policy.
- "The effects of policies are often indirect, diffuse and may take time to appear". As already mentioned, inference is made more difficult in a complex environment, which in turn implies that the evaluation of certain policies or experiments is challenging and might not readily translate into the creation of new

initiatives (or at least, not in the timeline required by the traditional policy process or government mandates, that is, two to four years).

While Hallsworth and his colleagues wrote this chapter with the UK policy cycle in mind, their insights apply to many other national contexts and are clearly pertinent for education systems.

The inadequacy of the standard policy cycle in a complex educational environment is further magnified by the fact that seemingly similar contexts may have very different dynamics and therefore be affected differently by the same policy. For instance, disclosing information about school performance might have very different impacts on a school that is thriving, as opposed to a school that struggles to attract well performing students. Whereas the traditional policy cycle usually operates on the back of a top-down framework of policy making, complexity instead requires policy cycles to operate closely enough to the local level so that policies can be tailored to idiosyncratic contexts, rather than follow a one-size-fits-all approach.

Two other elements contribute to making governance in complex environments challenging: a) the differences in objectives and time horizons of the various stakeholders and b) the informational challenges that arise as a function of the increased availability of data and information more generally. These will be discussed briefly in turn in the next sections.

Differences in objectives and time horizons

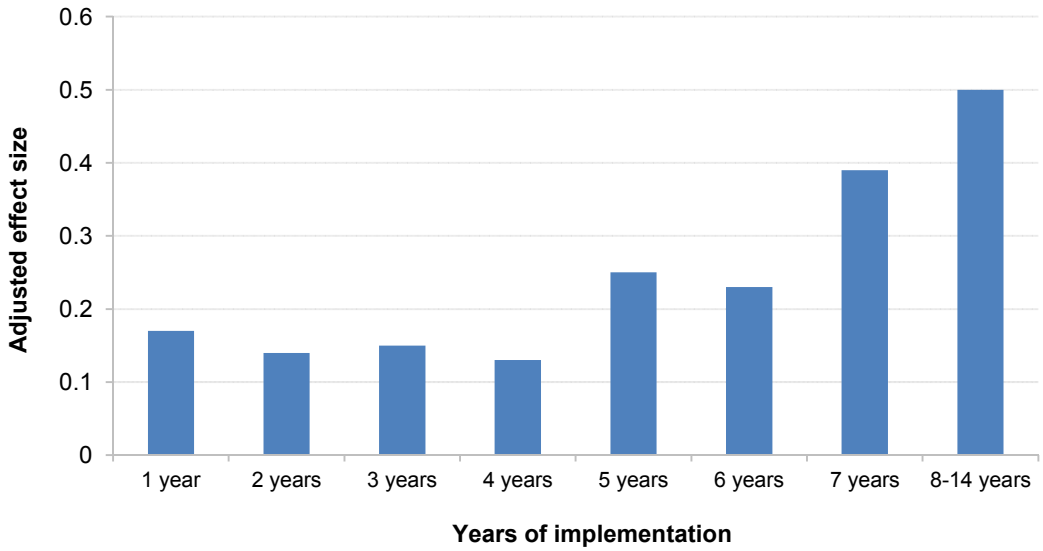
Two combined trends in the governance of education systems, and indeed many of the public sectors (e.g. health, justice, etc.) have jointly contributed to put stakeholders with very different perspectives at the heart of the decision making processes; on the one hand that of devolution of power towards lower more local levels of decision making, and on the other hand that of increased accountability and widened participation in the decision process. These trends have implied that education systems now tend to encompass more stakeholders than before: decision nodes have been more scattered both vertically within education professionals, and horizontally towards surrounding communities and other actors. The functioning of education systems has more than ever become the product of joint actions by students, educators, parents, school directors, civil servants and elected officials at various levels of government.

With this diversity of actors also comes diversity in expectations. For example, elected officials have to operate on shorter time scales than civil servants, teachers, parents and students. This can give rise to different policy preferences when choosing interventions, particularly as they relate to time and level of risk required. For example, quick-effect changes (e.g. providing students with electronic tablets) might become more appealing to elected officials as elections loom closer, while parents may favour longer-term less risky changes (e.g. reinforcing the teaching staff) and researchers may prefer more risky longer-term experiments (e.g. teaching a new reading method). Paradoxically, moving from appointed to elected officials as a way to increase local accountability in the education system, for instance in school boards, might result in an undesirable preference for more visible short-term solutions from those officials, given the requirements of the electoral cycle. The introduction of electoral accountability can also induce a succession of short-term reforms that may induce “reform fatigue” among the stakeholders impacted.

This preference for policies that yield effects in a shorter time span can sit uncomfortably in the realm of education, where policies may take a long time to take

effects. Indeed, one longitudinal model of comprehensive school reform suggests that the strongest effects are seen 8 to 14 years after a reform is begun, as shown in Figure 8.1 (Borman et al., 2003). This is due to a number of factors. In the realm of school choice, for example, Waslander et al. (2010) point out that reforms can take a long time to bear effects, both in terms of when the policy comes into practice and when parents act upon it. They also emphasise that time is not unidimensional but in fact interacts in highly complex ways with the actors and the context. In fact, short-term effects can be different from long-term effects because, using the school choice example again, certain types of parents might be quicker to react to the policy change than others. This implies that the kinds of impacts that are observed in the short term may be qualitatively different, both in type and intensity, than those that develop over a longer timespan. This is a particularly relevant observation when combined with the knowledge that policies do generally require quite some time to take full effect. Such conclusions show that it is sometimes necessary to leave to reforms a longer time span than the natural time span induced by electoral cycles.

Figure 8.1. Effect sizes of compulsory school reform in the years following implementation



Note: Effect sizes based on meta-analysis of studies pertaining to the impact of comprehensive school reform on student achievement in the United States.

Source: Borman et al. (2003).

This difference in time scales is compounded by a difference in objectives: officials, in particular elected officials, will favour finding "what works" while researchers may focus on understanding "why it works". Although ideally these two approaches should combine, in reality this is not always possible, in particular when facing budgetary constraints and thus the necessity of prioritising various issues.

Fractionalisation of decision making generates informational challenges

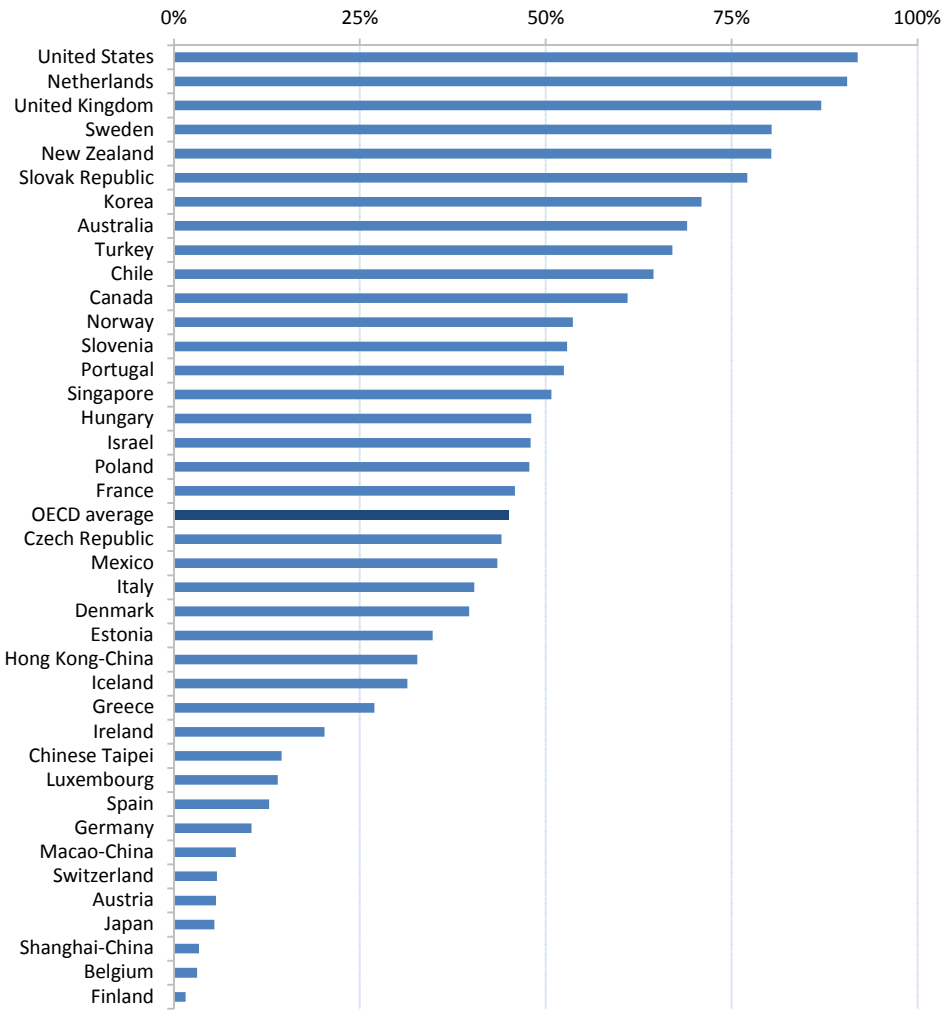
As already argued, education systems are becoming more complex through increasing devolution of decision making and the involvement of a wider range of stakeholders. They are also becoming more complex due to a greater availability of data on educational performance and other system factors that are relevant for the decision-making process.

The increase in the availability of information is one of the most dramatic transformations in our education systems in the last two decades, fuelled by two concurrent trends. The first is the rise of standardised tests (both national and international, for example PISA) and the resulting explosion of available evidence and greater emphasis on testing and assessment. The second is the increased access to information via the Internet and other technologies, which has enabled a multitude of actors to bring their own informed opinions to the discussion.

According to PISA 2012, on average across the OECD 45% of students are in schools whose principals report that achievement data are posted publicly. But this average belies a very wide distribution: in the United States, the Netherlands, the United Kingdom, Sweden and New Zealand, over 80% of students attend such schools. At the other end of the scale, less than four per cent of students attend schools where achievement data are posted publicly in Belgium, Finland and Shanghai (China) (OECD, 2014, see Figure 8.2).

Figure 8.2. Use of school achievement data for accountability purposes

Percentage of students in schools where achievement data are posted publicly



Note: Included in graph: all OECD countries and the top 5 PISA performers in mathematics.

Source: OECD, PISA 2012 Database, Table IV.4.13. http://dx.doi.org/10.1787/888932957346.

This abundance of data can pose challenges in an environment where different actors, with different needs, cultures and perspectives need to share information with each other. First of all, the availability of large amounts of data must not be confounded with having a full understanding of any given situation. Current data collections omit important (and potentially explanatory) variables on issues as diverse as student well-being, the role of non-cognitive skills in student achievement and motivation, teacher expectations, and of course a whole host of system-level variables. In complex environments these kinds of information can be as or more important in understanding interpersonal and institutional interactions than standard indicators on student achievement and teacher practice.

Secondly, even for standard measures, important information might also be only partially collected, or not systematically (for example, reasons underlying student drop-out or issues with teacher retention). It is true that the move towards more computerised environments and more accountable stakeholders has yielded a significant production of information that is now recorded by the majority of OECD countries (e.g. student achievement data, teacher assessment data, school budget data, school choices of parents etc.). However, the involvement of many stakeholders requires that the collected information is systematised in formats that make it easier to disseminate, analyse and compare nationally and internationally, and this is not always the case.

In addition, sharing information also comes with its own challenges. There might be few incentives for collected data to be shared widely, especially if there is a concern that it could be used in a negative manner (for example, in systems where there is strong competition for students between schools, the weaknesses of a particular school might be disguised or otherwise presented to avoid injuring the reputation of the school). Moreover, information might be hard to find, little publicised, or produced without thinking that it may be useful and hence not passed onwards to other actors in the system. The 2011 report of the Swedish National Agency for Education illustrates these problems; municipalities were shown to focus their attention on their rankings rather than performance, and used only a small share of the available data in its decision-making process (Swedish National Agency for Education [*Skolverket*], 2011).

On the other hand, too much information can obscure information pertinent to decision-making and/or render it unusable by its sheer magnitude. Loeb and Plank (2008) illustrate this danger with the California Education Code, which includes more than 100,000 articles and more than 2,000 pages. The abundance of information increases the difficulty with which stakeholders can learn about the existence of documents and then locate them. As O'Day (2002) points out, the abundance of information may even be counterproductive, as “teachers and schools may metaphorically and literally close the door on new information, shutting out the noise”. It also raises the question of how all the information can be gathered and maintained in a way that can be used by the other parties. In a complex environment with multiple active stakeholders, harnessing all this heterogeneous data and ensuring that it reaches those who need it becomes an important challenge.

There are two elephants in the room in this discussion. The first is that increased access to data (via the Internet and including media-friendly testing and assessment results which lend themselves so well to league tables and rankings) does not ensure that the quality of the information is consistently high. The Internet has effectively removed many of the established gatekeepers or quality controls that were traditionally put in place by research institutes and academic journals (Burns and Schuller, 2008). As they argue: “More information is available, yes, but is it good information? And is it presented

accurately and in an understandable fashion? Can the reader use it in a comprehensible and useful manner?" (p. 17).

The second elephant is that the *production* of data does not equate to its *use* (Fazekas and Burns, 2012). In all discussions of increasing the availability of data in order to increase transparency and accountability to a broader range of stakeholders, there is an underlying equity issue. In most countries upper middle-class and middle-class families (or parent(s) with higher education, higher professional positions and higher income) are the ones that are most aware of how to actively use the education system for their own interest and benefit (Taylor, 2009). They are also more likely to have the capacity to lobby and press for change in the educational system through policy and practice (van Zanten, 2003).

In practice this means that in many OECD countries middle-class parents are more likely than parents with lower socio-economic status to use school achievement and school performance data, when available, in order to place their child in the best-performing schools. If changing schools is not possible, middle and upper-class parents are more likely to demand (and successfully lobby for) change in the system. Parents with lower incomes (including, in many countries, high proportions of immigrant parents) are less likely to be aware of their rights regarding school choice and may often lack the capacity to use achievement and performance data. A similar argument can be made that some districts or municipalities might be more likely than others to fully use available data – perhaps those that care more about education quality, or those that have better capacity to analyse and interpret such data. These equity arguments are not trivial – indeed, any system motivated to provide full access to performance and achievement data in the name of transparency and efficiency cannot turn a blind eye to how and by whom those data are being used.

The availability of data per se then, is not a stand-alone solution to information asymmetries between stakeholders, and can in fact serve to increase the complexity involved in their interactions. PISA 2012 analyses make it clear that simply making school achievement data public is not correlated with better student outcomes. Indeed, among the top performers in PISA 2012, Shanghai, China (number 1 in math achievement) is at the very bottom of the Figure 8.2, and Singapore (number 2 in achievement) is just slightly above the OECD average in terms public availability of school data. Information can only lead to school improvement if it is relevant, available in adequate quantity, and properly interpreted (O'Day, 2002).

Experimentation in a complex environment

The governance challenges outlined above have led to the rise of strategies to deal with complexity and the dynamic nature of the system, its actors, and the data available to them. Policy experimentation aims to improve the system by explicitly testing new policy options and assessing which could be successfully generalised. In addition to education, policy experiments are used in a number of domains, such as development aid, healthcare, economic policy, etc. They rely on small-scale trials, and have as their objective not the immediate improvement of the system, but rather the discovery of what key factors would generate such improvements. In complex environments, decision processes must typically reconcile conflicting objectives among stakeholders. It is here that experimentation provides strong arguments which can provide solid evidence on which to constructively base the discussion. However, it is precisely in complex environments that such a task is made more difficult, for reasons detailed below.

What is policy experimentation?

There exists an abundant literature on policy experimentation, but for the purpose of this chapter, we rely on Heilmann's (2008b) definition of experimentation as "a purposeful and coordinated activity geared to producing novel policy options that are injected into official policymaking and then replicated on a larger scale". This definition entails three key components of experimentation, which will be discussed in turn below.

The first important element is that experimentation is seen as *a deliberate process*, and must therefore be distinguished from innovations emerging by chance or as unintended side effects to other deliberate processes. Experimentation presupposes that a problem has been identified, has been understood as relevant, and that one or more potential solutions have been suggested that need to be assessed. It therefore offers a systematic approach to dealing with a problem: 1) identifying the problem; 2) suggesting a potential solution, 3) trying that solution out, and 4) evaluating whether this solution was effective. The deliberateness of experimentation resides in the planning of the process, and in particular, in the ex-ante decisions of what is going to be implemented and how its effects will be measured. Experimentation therefore entails a certain methodological commitment, which makes it different both from spontaneous innovation and from reactive policy making.

The second important element is that experiments are *performed at small scale* that can then be fully rolled out if proven to be successful. Underpinning this attempt is the objective of assessing whether the experiment could suitably be generalised to a wider implementation. This implies that it is evaluated and the results of that evaluation are used in the decision of whether to scale up or not. Experimentation thus entails a process in which evaluation is an explicit step, by comparing outcomes between areas where the experiment took place and areas where it did not, in an attempt to assess its efficacy. In order to be able to evaluate its impact and effectiveness, it is necessary for the experiment to include a carefully designed and chosen control or comparison group that can act as a reference point against which to measure the results of the experiment (see Box 8.1). In that respect, experimentation is an approach that greatly differs from whole-system reform, which approaches change from a reform point of view at the scale of the whole system. This will be discussed more fully below.

Box 8.1. Treatment and control groups in evaluation

To correctly assess the effectiveness of a policy or a new practice, it is not sufficient to measure certain outcomes following the introduction such policy; it is also necessary to assess what proportion of these outcomes can be explained by the policy itself rather than confounding factors. If a new reading method is introduced that increases reading scores by 5% among students, it is important to understand whether, for instance, reading scores were not just improving among all students that year. If that were the case, the 5% improvement might give the false impression that the method was effective, whereas it was actually due to other factors that contributed to a general improvement in reading scores.

To isolate the effect of a new method or policy itself from other confounding factors, it is necessary to have a counterfactual, that is, to find out what would have happened if that method had not been introduced. The best way to construct such counterfactual is to have a treatment group that is subjected to the new policy or method and a control group that is not.

Box 8.1. Treatment and control groups in evaluation *(cont.)*

To make sure the comparison does not capture other effects, it is important the treatment and control groups are as similar as possible in all dimensions, except whether they were subjected to the new policy or not. Difference in outcomes between treatment and control groups can then be attributed solely to the new method or policy being evaluated. Assigning subjects (students, schools, districts) randomly between treatment and control groups can help ensure that groups do not differ in any systematic way.

The third element of this definition is that this exercise in assessment is used as *feedback towards policy making*. Experimentation is by essence supposed to produce new knowledge. First and foremost it reveals the potential effectiveness of the policy being tried out. It also reveals additional information, for instance regarding possible unintended consequences that were not part of the initial thinking on that policy, information regarding the feasibility of generalising the policy to a wider scale, how well the policy is received among the stakeholders, etc. As such, experimentation represents a form of evidence-based policy making, in the sense that it generates new information that can be used by policy makers to make educated decisions about the direction in which to steer education systems.

Policy experimentation is not inherently linked to any particular methodology, and, as long as the previous conditions are met, can be argued to encompass both the strict standards of randomised controlled trials (RCTs) [2] and the qualitative case studies that mark much of education research. Pilot programmes probably represent the most widely used type of policy experimentation today, although it must be highlighted that they qualify as experimentation if (and only if) they contain a comparison group and a proper evaluation phase to assess the results.

However, part of the problem with pilot programmes as experiments is that researchers or policy-makers have often chosen the participants and/or location of the programmes, usually for practical reasons. For example, researchers interested in the effectiveness of a new writing software might very well choose to pilot it in a set of schools nearby, both because that would facilitate their access to the schools for testing and because they are more likely to have personal connections to the headmasters and school boards who approve research proposals and agree to participate in them. Similarly, a new curriculum with an emphasis on sports activities might be tried out in a small number of schools facing student behaviour problems; with the expectation that this where such a programme is more likely to work. But that reasoning creates a problem down the line: even if the programmes prove successful, it is hard to know whether the intervention would also be beneficial in other schools, and whether it should be generalised.

Randomised controlled trials (RCTs) try to address this issue by assigning participants (students, classes, schools, teachers, etc.) randomly between a treatment group (where the new policy will be implemented) and a control group where conditions stay the same as before the experiment (see Box 8.1). By constructing two groups that are on average identical in their composition, any difference in the measured outcomes between the treatment and the control group can then likely be attributed to “treatment”, i.e. the policy being tried out.[3]

RCTs therefore offer the cleanest way to estimate the effect of a given intervention while being able to exclude confounding factors[4]. RCTs also allow for manipulating several variables at once, by setting up an experiment with several treatment groups in which experimenters vary some parameters, e.g. the length of the treatment, how many hours a week are devoted to it (intensity), etc. For instance, when Schwerdt et al. (2011) tested for the effectiveness of adult learning vouchers, they randomised the monetary value of vouchers, and could check whether their effectiveness depended on their amount (it did not). By being able to test for multiple treatments at once, RCTs are in this sense economical compared to pilot programmes which, by essence, can only test for one set of parameters.

But such clarity and testing power comes at a price: RCTs typically require samples (of students, classes, schools, etc.) big enough that it is possible to attribute difference of outcomes between treatment and control to the intervention rather than to chance. In addition, while they may tell you what works (or does not work), they do not generally answer the question of *how* and *why* something works. Understanding those elements requires additional evidence often gleaned through other methodologies (OECD, 2007).

Sometimes, a change or reform occurs outside the scope of a carefully designed experiment. While prior design is always preferable from an inference point of view, it might in some cases still be possible to use quasi-experimental methods to estimate *ex-post* the impact of the change. In particular, when there are reasons to believe that the change affected subjects at random[5], the same comparison analysis as in RCTs is still valid.

In the field of education, regression discontinuity design (RDD) is often used to estimate the effect of certain factors in quasi-experimental settings, and is even recommended when random assignment is not possible. The technique compares subjects on one side and the other of a boundary that determines to what policy they are subjected (the boundary creates a discontinuity in policy). The boundary can be a physical or geographical boundary (e.g. district limits), a threshold grade, the birth month or year, etc. For instance, van der Klaauw (2002) compares students with scores closely under and above the threshold to qualify for financial aid, in order to estimate the returns to college attendance. The idea of RDD is to take a narrow band of subjects around the threshold, so that the groups are similar and can be treated as if they had been randomly assigned above or below. In general, these quasi-experimental approaches are harder to control, but research among education stakeholders using these designs can be encouraged in order to gather helpful information without necessarily having to incur heavy costs that can sometimes affect pilot programmes or randomised experiments.

Although experimentation requires the collection and analysis of quantitative data to evaluate the treatment being considered, it does not exclude other more qualitative tools. Experimentation does not occur in a bubble. Qualitative research such as surveys and case studies can yield important insights in a number of ways. Qualitative techniques can be used to help select hypotheses to be tested and refine the design of the experiment. After and during an experiment, qualitative techniques can help tease out various mechanisms and uncover potential explanatory factors and secondary effects.

Inference in a complex environment

In order to improve systems, researchers have argued that changes should be implemented in a way that could foster learning opportunities about what works and what does not (Campbell, 1969). The ability to learn from a particular experiment is inherently

linked to the quality of its original design. However designing and conducting research that yields this information is not always easy: in particular, a randomised sample might be difficult to obtain in many fields, including education (Campbell and Stanley, 1963). This is true for a number of reasons, with the most pertinent one being that participation in educational experiments might be hard to enforce. In such a situation, attrition could lead to a smaller sample, a biased sample[6], or simply to the experiment being cancelled (see Box 8.2).

Box 8.2. The problem of attrition in education experimentation

Attrition refers to the fact that some entities (schools, classrooms, students) may drop out before the experiment is scheduled to finish. This can be problematic because it is harder to assess differences between treatment and control groups once the sample size becomes too small. In addition, there is a possibility that attrition is not random, but rather biased in a way that would affect the comparability of the treatment and control groups. For instance, perhaps wealthier parents would choose to relocate their children to non-participating schools, thereby leaving on average poorer students in the one group. This can then give a false picture of treatment effectiveness. Although it is possible to compare the composition of both groups after the fact, selective attrition cannot be corrected for in statistical analyses.

In education, this can be particularly problematic because participation in experiments is often voluntary. Since the intervention is new, stakeholders might not always be willing to participate. Offering incentives (monetary or otherwise) to participate in the experiment has been suggested as one way to reduce attrition or outright refusal to participate in an experiment. The downside is that it increases the cost of the experiment. It might also serve to select participants that are more responsive to such incentives, and thus not be representative of the entire population.[7]

Attrition raises a key issue for experimenters as *inference*, the task of learning from the data generated by experiments, can only be done when it is possible to compare the outcome of the experiment with a reasonable counterfactual. In a complex environment, inference of cause and effect becomes, well, complex. Even randomised controlled trials (RCTs) provide estimates for what would be the average effect of a certain policy, holding everything else constant. They identify what would happen on average if such a reform were to be extended to a similar group. However, the notion of an average effect might not be relevant in an environment that is characterised by important idiosyncrasies and non-linear dynamics. For instance, an experiment involving the provision of free textbooks for students might have very different effects depending on whether it is applied in a wealthy district or in a district where households face harsher budget restrictions. Effects can also be non-linear when a policy requires a critical size to be effective, for example with vaccination campaigns, which are ineffective unless they reach a significant share of the population. It may be important to understand how a potential policy might affect various contexts differently and if possible to find potential explanations for such variation.

Another complicating issue for experimentation, particularly salient in the realm of education, is that individuals may exhibit different behaviours if they know they are part of an experiment – a phenomenon sometimes called observer effect or Hawthorne effect. When this takes place, the whole subject pool (both control and treatment group) becomes systematically different from the rest of the population.[8] Left ignored, such observer effect might give a distorted picture of what would happen if the experiment

were to be generalised. In practice, this effect is more likely to take place in situations where individuals are expected to provide some efforts towards a task. For instance, in their experiment on low-stake feedback to teachers in Andhra Pradesh (India), Muralidharan and Sundararaman (2010) found that teachers altered their classroom behaviour even when they were assigned to the control group. While it is theoretically possible to remove this observer effect by comparing with individuals who are not part of the experiment, it is not always feasible to do so, for practical or ethical reasons.

Lastly, inferring the effect of a policy is accomplished by isolating it from confounding changes that occur at the same time. This *ceteris paribus* assumption, i.e. that all other things (than the policy) are held constant, might not be a trivial one to make in a complex environment, characterised as it is by fluctuating contexts and tipping points. The existence of feedback loops, especially reinforcing ones (usually termed positive feedback), may generate systems that exhibit several equilibria between which a system might oscillate, or on the contrary be permanently stuck. The capacity of the system to absorb changes, or on the contrary to exacerbate them, could then generate very different responses to the same input. When such path dependence[9] exists, a given reform might yield very different outcomes even when applied to seemingly similar systems. Recent research in the Netherlands (van Twist et al., 2013) has demonstrated that for example the assignment of the label "very weak" to a school can elicit a positive response from one school and a negative response from another, depending on the local context, history and staffing situation at the school; in self-reinforcing processes. Such sensitivity renders the analysis of the causes of success or failures for any particular policy intervention more complicated.

Bottom up experimentation is important but not sufficient

The salience of local discrepancies between schools or districts may suggest that bottom up initiatives are the best way to deal with educational challenges. If schools vary greatly from one another, then it is possible that solutions emerging from each school would be better suited to tackle their respective problems. However, in a complex system, bottom up initiatives cannot be scaled up to the broader system without at least some level of centralised discussion. As such in such a context, experimentation offers a systematic approach to this discussion.

The Dutch example outlined above (van Twist et al., 2013) illustrates a key insight. Namely, when a system exhibits self-reinforcing dynamics[10], some elements will benefit from virtuous cycles (success breeds success), while others will be caught in vicious circles (where difficulties bring about further difficulties). The consequence of such differences in local dynamics means that some heterogeneity will exist and persist in the system. Small initial differences in local contexts can therefore be exacerbated[11], creating a situation in which important discrepancies between schools or districts can persist and become hard to mitigate. Drawing from the Dutch GCES case study (van Twist et al., 2013), after receiving the "very weak" rating, some schools were facing a vicious circle of attrition among their students and staff, triggered by negative assessments by the Inspectorate; other schools experienced a virtuous cycle where parents would trust the school to perform adequately, which in turn led to a cooperative culture in which school staff and parents could work together. These discrepancies imply that careful attention must be paid to the particularities of each educational context.

The importance of context in education may suggest that local solutions might be a useful way to identify new, broader, policy options. However, scaling up local solutions

is hampered by the specificity of particular success stories. Generalising the success of one initiative to another school, district or wider area may generate very different results from the ones initially expected. As a result, although bottom up initiatives are an important way to generate innovative ideas from the field and suggest new policy options, decisions about rolling them out on a larger scale must be taken within the context of a centralised discussion about their potential for generalisation and sustainability on a larger scale.

Heilmann's analysis of policy experimentation in China can shed an interesting light on the experimentation process in a geographically large and culturally diverse context. He argues that China's economic success can be partially explained by the so-called "point-to-surface technique", an approach to experimentation which "gives room to local officials to develop models on their own, while ultimate control over confirming, revising, terminating and spreading model experiments rests with top-level decision-makers" (Heilmann, 2008a). The approach lets individual decision-makers (points) experiment and implement new ideas, which are then reported higher-up in the hierarchy, and if approved, then generalised to the rest of the constituency or country (surface).

This freedom in experimentation at the local level, later sanctioned or discarded by higher levels of government to fit within the central government's objectives, allows responding to local specificities of each economic area. However, Heilmann's analysis highlights the importance of having information and results from (possibly spontaneous) local initiatives and experiments disseminated to higher levels of decision making, which then decide whether to implement the experiment on a wider scale. Although the specificities of China's governance process do not map well on many OECD countries, it is an intriguing example of one approach to addressing what has been labelled the "tight but loose" issue (Thompson and Wiliam, 2008, cited in Hopfenbeck et al., 2013). That is, an approach to implementing reform (or in this case system level change) that keeps in mind core central principles (tight) while leaving room to implement these principles according to an individual's or region's initiative (loose).

It is a similar intent that was behind the change made in 2005 by Flanders in Belgium, which set up a three-year period during which schools were allowed try out temporary projects and experiments in pre-specified list of domains. As a result, individual schools applied for authorisation in order to experiment with things such as curriculum, extra help to non-native speakers, teacher autonomy, contractual arrangements, peer-review of teachers, etc. The idea was to allow schools to experiment, and later on take stock of what was successful or not, with a view to feed those insights into policy-making. That said, it is debatable whether such a non-systematic approach could yield learning robust to other contexts. The evaluation report was limited when drawing general conclusions from those varied experiments (Ministerie van de Vlaamse Gemeenschap, 2009).

Indeed, despite the intentional element to the inductive approach to experimentation, illustrated by the Chinese and Flemish examples, important subjective judgements are still made as to what could be successfully scaled up. As a result, while using unfettered local experimentation can stimulate the emergence of new ideas, this type of analysis still relies on post-hoc selection of initiatives based on the subjective assessment that they are transferrable and that they comply with grander (often politically motivated) objectives.

The use of carefully designed policy experimentation can offer a more harmonised framework in which policies can be tried, at a smaller scale, and in various contexts, as it can substitute for a more systematic approach to finding and scaling up novel policy solutions. In paying attention to the way the experimentation is deployed, it can also help

identify what contextual factors can affect success or failure, an important element to take into account when determining whether, where, and how the experiment can be successfully scaled up. Experimentation therefore attempts to remove the subjective part of the judgment, and replace it instead with a more systematic approach that identifies success criteria ex-ante.

Experimentation in the debate between whole-system change and gradual reforms

In complex environments, all agents are embedded in a network in which they interact with each other. This means that ultimately, the success or failure of certain educational practices can be only assessed by taking all stakeholders into account. There are two contrasting ways to approach this challenge. One approach emphasises a whole-system approach to change in order to tackle concomitant issues at once; the other approach emphasises instead the necessity of small changes in order to more clearly identify the effects of the intervention and bring stakeholders on board. Experimentation falls into that later category. What follows is an overview of the two perspectives.

The whole-system approach argues that change should occur at a greater scale in order to capture all relevant nodes of a system. Targeting change too narrowly might be unsuccessful, because it affects only a small part entangled in a much bigger network. This whole-system approach is for example suggested by Fullan (2009), who welcomed the emergence of numerous intentional system-level reforms since the mid-1990s and advocated a multilevel approach to change that encompasses schools, regional subdivisions/districts and the central government. This is also the approach suggested by Mason (2008), who argues that feedback loops endow complex systems with a significant "inertial momentum". Consequently, "what it might take to change a school's inertial momentum from an ethos of failure is a massive and sustained intervention at every possible level until the phenomenon of learning excellence emerges from this new set of interactions among these new factors, and sustains itself autocatalytically[12]."

This echoes arguments made for instance by Lipton and Sachs (1990) about the transition of Eastern Europe to market economies in the 1990s. They argued that states should take advantage of their *état de grâce* to implement a "big bang" approach to (economic) reforms, sometimes referred to as "shock therapy". In the realm of education, recent experience has for instance shown that the publication of PISA results has, for some countries, constituted a wake-up call that prompts public discussion on education and yields momentum for sweeping reform.

While a whole-system approach might be a successful solution to facilitate change in a complex environment, it might be problematic from a learning point of view. The goal of experimentation is fundamentally that of learning: experiments are designed to assess the effectiveness of new policies or practices in order to offer new viable policy options. The whole-system approach to reform runs counter that objective by offering a blanket approach to change.

Popper (1957) argues against whole scale change on this epistemological ground and instead favours a "piecemeal" approach, sometimes referred to in the literature as "gradualism". Popper argues that a piecemeal approach "permits repeated experiments and continuous readjustments". Since learning occurs through trial and error, Popper emphasises the importance of being able to identify the causes of success or failure of a change. According to Popper, as the holistic approach does not offer a counterfactual, it is not possible to identify whether successes or failures are due to the policy, or whether

outside elements confound the results. It then becomes impossible to eliminate a bad policy and replace it with something better. Instead, he favours a more gradual approach that operates on the basis of successive incremental changes. It then becomes possible to assess success or failure to each of these steps, and to have a clearer picture of what works and what does not. This is also the point made by Campbell and Stanley (1963), who argues that, as much as possible, reforms should be rolled out in a way that can lead to proper evaluation. For instance, when random assignment is not an option, they offer staggered interventions as a possible alternative to be able to compare groups in a meaningful way.

A further argument in favour of gradualism and experimentation can be made on more pragmatic grounds: gradual changes are easier to accept and therefore to implement. This is in essence the point made by Dewatripont and Roland (1995), for two reasons. First, gradualism can help build support for reforms over time: initial successes can form a basis on which to build popular support for subsequent changes. Second, gradualism ensures that if the initial reforms do not work as intended, the cost of reversal is lower than if a larger collection of reforms had been implemented.

These two opposed views highlight an implicit tension in complex systems, between learning and realism. On the one hand, understanding whether a policy would yield the desired effects typically requires a smaller-scale well-designed experiment, which allows for rigorous comparison with (near) counterfactual settings. This enables a more precise evaluation of the policy option being considered, the effectiveness of which can therefore be assessed in order to determine its suitability for generalisation. On the other hand, when operating on the basis of narrow interventions one risks missing the effects of certain reforms that require several changes to be made simultaneously or that require a critical size to be effective. In such a case, the interconnectedness of the system constitutes a threat to the external validity of the experiment: the change might have different effects if it is scaled up. One must then adopt a more holistic approach that takes that interconnectedness into account if one aims to understand the effects a policy would actually have if once scaled up. And of course, there is always the question of what is feasible given existing political and budgetary constraints.

This tension is echoed by similar debates in other fields. For instance researchers and policy makers working on developing countries have debated whether economic interventions should take a big-package approach, in which multiple objectives are pursued at one time (see for instance the Millennium Villages Project[13]), or a more targeted approach that enables more learning (Banerjee and Duflo, 2011). This debate in economic policy parallels the debate between whole-system approaches versus piecemeal changes, in that whole-system tries to generate a big change momentum by addressing many problems at once, but presupposes that the solutions they offer are the correct ones, whereas targeted change takes a more cautious and agnostic approach, but can only work on smaller well-identified problems.

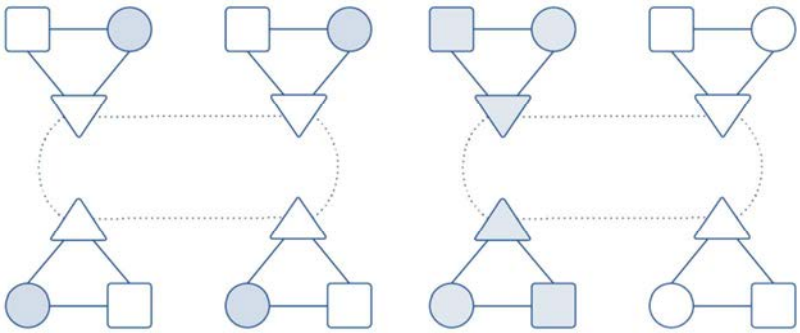
That tension between the possibility of experimentation and the accuracy with which it can be done (taking complexity and interconnectedness into account) can never be fully resolved. Since the goal of experimentation is to learn, experimenters and designers ought to err on the side of narrow but well-identified experiments. However, a good balance can sometimes be struck by designing experiments that operate on units of significant scales, or ecosystems. This is detailed below.

Moving towards ecosystem experimentation

The challenge of complexity is acknowledging that while no perfect solution exists, it is possible to take small concrete steps to make a difference. Policy experimentation can be made more useful by choosing an appropriate scale and design, called *ecosystem experimentation*; by adopting the policy cycle to reflect the dynamic nature and the intricacy of education systems; and by ensuring that input from stakeholders is also matched by a culture of constructive criticism within the system that can identify successes and failures.

Embracing ecosystem experimentation involves moving from horizontal experimentation, where experiments focus on a certain type of node (e.g. changing the reading teaching method in all schools), towards experiments are conceived as focussing on self-contained parts of the systems (i.e. natural ecosystems).

Figure 8.3. Horizontal experimentation (left) versus Ecosystem experimentation (right)



Note: Dotted lines denote weak links; solid lines denote strong links between elements of the system.

Figure 8.3 illustrates the distinction between those two concepts. On the left panel, the horizontal experiment targets all nodes of a certain type (in this case, the circles in grey, which may for example represent schools[14]). On the right panel, the experiment focuses on self-contained subsets of the system, for instance the networks of all stakeholders within a small set of given districts. The idea behind targeting self-contained subsets is to identify pockets of the system in which the links and influences within the subsystem are much stronger than links with other parts of the system. In the figure above, these pockets are represented by a triad circle-square-triangle. Note that the interactions are strong within a triad (solid lines), while interactions *between* triads are weak (dotted lines).

However, the reality of today's education systems is that no part of the system is completely isolated from the others. Instead, many outcomes (in the broad sense) result from the interactions of many actors. Let's consider the example of healthy dieting, which has been identified as a crucial element of education (OECD, 2014). Concerns about increasing obesity have emphasised that students' eating habits must involve parents, schools and teachers. From an education perspective, an experiment trying to change students' eating habits might try a different layout of cafeterias in a number of randomly selected schools, a horizontal experiment. But parents constitute an important factor in children's diet, and only changing food available in school might not suffice. Instead, school menus and parental food choices might reinforce each other (for better or worse).

When such complementarities exist, it might be more fruitful to randomize at a higher level (e.g. assign whole towns to either treatment or control) and involve both cafeterias *and* parents in the experiment (e.g. by providing parents with information, workshops, etc.). Such an ecosystem experiment would be able to capture the joint effect of school food and parental influence (without, unfortunately, being able to distinguish their respective effect).

When trying to target ecosystems, a new question emerges: where to draw the boundaries of any given ecosystem? In an intricate environment it might seem *prima facie* difficult to isolate subsystems that are self-contained. However, research in ecology shows that in complex networks, each node is mostly affected by other nodes within two or three degrees of separation (see for instance Williams et al., 2002). Applied to the educational paradigm, this means that it is often possible to isolate a relatively separate subpart of the system, in which nodes strongly influence each other but are not too affected by the other part of the networks[15]. In a country where education is managed at the district (or municipality) level, these could represent independent subparts of the system. In metropolitan areas, neighbourhoods might constitute such subparts. The definition and boundaries of such ecosystems will of course depend on the nature of outcome considered. One example of this multi-stakeholder experiment is the work of the Harlem Children's Zone (HCZ), a wide-ranging programme set in Harlem, New York.[16] From its inception, the objective of the HCZ programme has been to tackle a number of issues simultaneously, both inside and outside schools, in order to create a tipping point towards a positive learning environment. For eligible communities, the intervention combines two elements. First, children can benefit from a pipeline of charter schools, from parenting classes, followed by early childhood interventions, all the way to high schools, and support for college entry. The schools in the programme also offer extra-hours, social activities, healthy foods, sport programmes, as well as tutelage in non-cognitive and employability skills. In parallel to that pipeline, the second tool of the HCZ programme is to invest in community programmes, such as cooperatives, health and food education to families, healthcare initiatives, community centres and foster care prevention. The programme therefore combines a move towards charter schools, with investments in the surrounding communities, in order to involve a wide range of education stakeholders. The HCZ has been estimated to be very successful at bridging the racial achievement gap in the targeted area (Dobbie and Fryer, 2011.

Perhaps the most important consideration when choosing the scale at which the experiment takes place, and the level at which to randomise, is that the best scale often depends on the relevant type of interaction between agents being studied, and therefore on the question being analysed. In education, one of the most common concerns is therefore whether to randomise at the student level, at the teacher level, at the classroom level or at the school level.[17] Some interventions might even want to consider randomisation at the school district level. In general the statistical analysis is made easier if the randomization takes place as close as possible to the final level of analysis. So, when investigating the effect of a given intervention on student's scores, it would seem easier to allocate students randomly between treatment and control groups, for example.

However, in certain experiments where there are significant interactions between students, student-level randomisation might not be appropriate. Consider for instance a hypothetical experiment that would look at the effect of providing more information about future careers on students' motivation. Since peer effects could be important factors that affect educational and professional aspirations, it might make sense to account for peer connections between students when designing the experiment. Randomising at the

student level would lead to classes comprised of students from both the treatment and control groups. If students within classes interact with each other extensively, it is likely that information provided to students of the treatment group would leak to students of the control group; or that behaviours from one group could affect behaviours of the other group's members. This "contamination" would likely make outcomes of treatment and control groups more similar; as a result, it may give experimenters the – possibly wrong – impression that the experiment did not have any effect. This is a situation where it might make more sense to allocate whole classrooms, or even whole schools between treatment and control, in order to minimise the spread of information, and other contaminations effects, between treatment and control. In general, experiments that involve information campaigns are more likely to suffer from issues of contamination, and are probably better randomised at some reasonably high level, such as school- or district-level. It is however useful to remember that this comes at the cost of reduced statistical power.

As mentioned, another question that might arise during the design of certain education experiments is whether to randomize at the teacher or at the classroom level.[18] Consider the example of a new teaching method, say in mathematics. To assess its efficacy, it might useful to train some teachers so that they use that new method in class, while others stick with the traditional method. But doing so creates the risk, although improbable, that the teachers selected are not representative of the other teachers; if this were the case, then assessing the efficacy of the new method based on those teachers would give a wrong picture of what would happen if the method were extended to other teachers. A potential solution to remove the "teacher effect" is ask trained teachers to use only in some of their classes (chosen at random), and compare outcomes in classrooms where the technique is used and classrooms where it is not. For this setup to be valid, it requires that teachers can "turn off" the new method, and teach in two different ways, without the new method contaminating the way they teach traditionally. If the experiment were instead an information campaign towards teachers, e.g. reminding them about gender issues among students, this might be problematic, as it is difficult to imagine that teachers could "turn off" this newly-acquired information. In that case, classroom-level randomization would likely suffer from contamination through the fact that same teachers teach different classes. It would in such case be more appropriate to randomize at the teacher-level. Again, the scale to choose for the experiment, and in essence, the appropriate definition of what constitute an ecosystem, depends on the question being examined.

Paradoxically, while complex systems might be where ecosystem experimentation is the most necessary, it will also be where they are the hardest to implement in practice. Since randomised designs aim at obtaining a precise evaluation of an experiment, they are particularly vulnerable to attrition. What if the units selected for the policy trial would rather not participate? In many countries, ministries might not have the power to enforce enrolment in the program, nor might they be able to prevent opting out while the process is underway (see Box 8.2 on attrition above). For such a design to function, it is imperative that excellent communication and promotion to all stakeholders involved are present. Such communication must clearly explain the content of the approach, how everyone will be affected, and the question the experiment is trying to answer. A key benefit of involving all stakeholders at once is that it avoids the pitfall of ex-ante singling out one stakeholder as a bottleneck towards educational achievement.

It is also important for all stakeholders to understand that their participation is crucial for the experiment, regardless of the outcome. Although an experiment could reveal an amazing improvement in the variable under study, it could also reveal that there is little,

or no, effect of the intervention. Although not as "sexy" as observing strong improvement, knowing when something does not work – or works but only to a limited extent – is also crucial to making policy[19]. In times of limited resources, it is extremely useful to know what will give the greatest return on investment, and this information can only be learned through trying things out – ideally, through experimentation.

Adapting the policy cycle

How would this work in practice? The premise of the complexity approach to education is that success or failure of schools depends on many interlinked stakeholders who operate in a fluctuating environment. Only by involving all relevant stakeholders at once is it possible to generate a switch from one type of equilibrium to another, in order to overcome what Mason (2008) calls "inertial momentum". A key tool for ensuring the involvement and participation of all stakeholders is to move towards a more "open source" input approach to experimentation process. By widening inputs for ideas and comments, notably through more integrated IT systems (Loeb and Plank, 2008), it is possible to make sure that all stakeholders can participate in the identification of problems, have their say in their prioritisation and can suggest potential solutions (Box 8.3).

Box 8.3. Technologies in education governance

Information technologies have affected the way teaching takes place, and also the way schools, teachers, students, administrators and local authorities interact with each other. In addition to using information and communication technologies (ICT) to manage scheduling, staff, students and grades internally, many schools have also turned to ICTs to facilitate contact with parents and the wider community and promote more inclusive governance and accountability. For instance, the Eudora school district in Kansas (USA) has set up explicit policies and guidelines to facilitate online presence by teachers, school staff, and teacher-parent contacts, to both share everyday activities and to inform and deal with problematic situations (Lepi, 2013).

In open data initiatives, cities or governments decide to make their data freely available. The data is designed to be accessible to local authorities, schools and headmasters, and parent and community members. Relatively recent, these initiatives are becoming widespread: examples in education include the MySchools website in Australia, the New York City Department of Education's School Choice Design Challenge, Sweden's Open Comparisons website, and SIMCE in Chile.

Frequently inspired by issues of school choice and tracking student achievement, these trends highlight the greater involvement of parents and the community in the governance of education systems. This involvement can be formally structured and solicited, as in the examples above. But it can also be informal and rely on the power of ICTs and social media to unite actors brought together by a shared willingness to improve education in their communities.

Policy experiments can provide essential evidence when considering whether an evaluated policy should be rolled out on a greater scale. In particular, if the experiment is properly designed, its results should be used to inform future policy making and feed reflexion on the causes of success or failures. Experiments can and should constitute important elements to hold policy makers accountable: they should be able to justify costly wide-scale reforms using sound knowledge, including experiment-based knowledge. The analysis and dissemination of the results are not designed to blame

specific individuals but rather as necessary feedback for a better understanding of what works, what does not, and how this new knowledge can be taken further. It is a necessary condition for establishing a trial-and-error system of successive experiments in which constant learning is possible. The knowledge for all participants that they are ultimately contributing – possibly in an indirect and implicit way – to the improvement of future generations' education systems and practices, can be a strong motivating factor.

The process of experimentation should also be able to adapt to changing conditions, by continuously adjusting its methods and processes. As such it may be possible to design experiments that undergo several iterations to address successive as a form of "experiment package" designed to address related questions at once (Cobb et al., 2003) or to successively narrow down the factors of interest that may help explain specific outcome dimensions (Box, 1999). However, this must be done with caution. A good balance needs to be achieved between the ongoing tweaking required to adjust to feedback loops and the need for experiments to be sustained for long enough that their effects have time to set in. Minor changes can still be accommodated reasonably easily in the experimentation design without nullifying all previous data, but attention should be devoted to ensure that experiments are carried on long enough for meaningful data to be gathered.

Conclusion

This chapter analysed the impact of complexity on policy experimentation in educational systems and suggested potential ways that it can be harnessed. The difficulty of experimentation is that it requires the involvement of many interdependent stakeholders with different cultures, motivations and time horizons, in a careful exercise of co-ordination towards a common goal of identifying policies that work and policies that do not. Governing in such an environment poses many challenges.

The complexity approach acknowledges the interdependence of all stakeholders in education systems, and sees managing this interdependence as a key factor in educational achievement. In such a context, a linear approach to policy-making will not work adequately. The process of identifying issues and suggesting novel policy options to solve them must therefore rely on experimentation, which can be a source of solutions that can take account of local contexts while preserving a systematic approach to solving issues.

When turning an idea into an actual experiment, designing the implementation of the experiment adequately must take this complexity into account. Experiments must strike an appropriate balance between two antagonistic objectives: on the one hand, involving all the stakeholders that could affect the experiment's success or failure, which requires operating at a wide enough scale; on the other hand, designing an experiment that can be rigorously evaluated, which requires experimenting with targeted changes. Although these objectives can never be fully brought together, the chapter suggests *ecosystem experimentation* as a potential solution: identifying self-contained parts of educational systems, such as districts, and randomising their allocation between treatment and control.

Notes

1. For a more detailed analysis of complexity and its impact on educational governance and reform, see Snyder (2013).
2. Such experiments involve a wider sample, divided at random between those who are allocated to the treatment (the policy being tried out), and those who serve as a control group. This design ensures maximum comparability of the groups in order to better assess the effects of the treatment. In the field of education, this was famously pioneered by experiments such as the Tennessee STAR (Student-Teacher Achievement Ratio) experiment in the United States, to assess the effects of class size on student learning, or more recently the extensive work by Fryer (2011) on what types of incentives affect student outcomes.
3. It is possible to construct experiments that allow for the comparison of multiple policies (one at a time), by constructing as many treatment groups as there are of policies to be tried out, while retaining one control group. This of course requires a bigger pool of participants.
4. The crucial idea behind RCTs (and in general any statistical analysis trying to estimate the causal effect of one variable on another) is that the variation in the causal factor should be unrelated to other factors that could affect the outcome. If one allocates subjects randomly, by definition the causal factor (being in the treatment instead of the control group) will be unrelated to other characteristics that might affect the results (e.g. age, ability, socio-economic background, etc.). It then becomes possible to establish the causal effect of the treatment on the outcome (e.g. student literacy rate), while excluding possible confounding factors.
5. Researchers talk about a "natural experiment", as if nature had allocated at random subjects between treatment and control groups.
6. If attrition is correlated with certain individual characteristics, the sample will not be representative anymore.
7. Despite this selectivity problem, policy makers may sometimes be interested in the effect of *offering* a certain programme, where take-up is ultimately voluntary.
8. A good overview of problems of attrition and biased subject pools in field experiments can be found in Harrisson and List (2004).
9. Path dependence is the property of systems in which the final outcome depends extensively on previous conditions, so that a small change in previous conditions might yield large changes in outcomes. A tennis ball placed on the top of a pitched roof exhibits path-dependence because a small push in one direction or the other determines on which side of the house the ball will fall. A counter-example is a marble dropped in a convex bowl, which always ends up at the bottom of the bowl regardless of the initial starting point. Systems that exhibit positive reinforcement loops can easily exhibit path dependence (Page, 2006). One education example is when reputation can help achieve better performance. For instance, if parents can choose where to enrol their children, a school that has a better reputation can attract more students and is therefore able to select the highest achievers more easily. Conversely, a school that is expected to perform poorly will fail to attract good students, and as a result confirm the initial prediction, a self-fulfilling prophecy.

10. These self-reinforcing dynamics are sometimes referred to as positive feedback loops. They exist when the more a phenomenon takes place, the more it feeds itself.
11. This property is also known as path dependence.
12. Autocatalytic reactions are chemical reactions that produce the catalytic compound which enables the reaction to take place. This means that once they start taking place, they fuel themselves; Mason uses this metaphor to talk about dynamic changes in education systems that can sustain themselves once they have reached a critical threshold.
13. See www.millenniumvillages.org/ (accessed 16.12.2015); see also Sachs (2006).
14. Instead of certain entities, such as schools, experiments can also target processes, such as teaching methods, or reporting workflow in schools, etc. The argument remains identical: a consistent set of processes should be targeted at once to have a more comprehensive understanding of the total effect of the experiment. This does not preclude the use of randomization and control groups.
15. For examples of how this identification might work in practice, see Snyder (2013).
16. See Harlem Children's Zone project website, www.hcz.org
17. Randomising at the student level means allocating individual student randomly between treatment and control group. Randomising at the teacher level means that teachers are individually (and randomly) allocated between treatment and control; if their students are part of the experiment, then all students of a given teacher should be allocated to the same group (treatment or control) as that teacher. And so on for school- or district-level randomization.
18. This echoes concern in clinical medicine about whether to randomize at the patient level or at the practitioner's level.
19. Clearly this is a difficult message from a political point of view. Communication with the stakeholders should thus make clear that experimentation is in fact also about efficiency, in the sense that understanding the effectiveness of a suggested policy is good value for money, if even the result is neutral or negative.

References

Angrist, J.D. and V. Lavy (1999), "Using maimonides' rule to estimate the effect of class size on scholastic achievement", *The Quarterly Journal of Economics*, Vol. 114/2, pp. 533-575, http://dx.doi.org/10.1162/003355399556061.

Banerjee, A. and E. Duflo (2011), *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, Public Affairs, New York.

Borman, G.D. et al. (2003), "Comprehensive school reform and achievement: A meta-analysis", Review of Educational Research, Vol. 73/2, pp. 125-230, www.rer.sagepub.com/cgi/content/abstract/73/2/125.

Box, G.E. (1999), "Statistics as a catalyst to learning by scientific method, part II – a discussion", *Journal of Quality Technology*, Vol. 31/1, pp. 16-29.

Brown, A.L. (1992), "Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings", *The Journal of the Learning Sciences*, Vol. 2/2, pp. 141-178.

Brown, L. and S.P. Osborne (2013), "Risk and innovation", *Public Management Review,* Vol. 15/ 2, pp. 186-208, http://dx.doi.org/10.1080/14719037.2012.707681.

Burns, T. and T. Schuller (2008), "Evidence-informed policy in education: New opportunities and challenges", in R. Desjardins and K. Rubensen (eds.), *Research of Policy and Research for Policy in an Era of Transnational Education Policy Making*, Aarhaus University Press, Aarhaus.

Byrne, D. (1998), *Complexity Theory and the Social Sciences: An Introduction*, Routledge, London.

Campbell, D.T. (1969), "Reforms as experiments", *American Psychologist*, Vol. 24/,4, pp. 409-429.

Campbell, D.T. and J.C. Stanley (1963), "Experimental and quasi-experimental for research on teaching", in N.L. Gage (ed.), *Handbook of Research on Teaching*, Rand Mc Nally & Co., Chicago.

Chingos, M. M. and G. J. Whitehurst (2011), *Class Size: What Research Says and What it Means for State Policy,* Brookings Institution, Brown Center on Education Policy, Washington, DC, www.brookings.edu/research/papers/2011/05/11-class-size-whitehurst-chingos.

Clay, E.J. and B. Schaffer (1984), *Room for Manoeuvre: An Exploration of Public Policy Planning in Agricultural and Rural Development,* Heinemann, London.

Cobb, P. et al. (2003), "Design experiments in educational research", *Educational Researcher*, Vol. 32/1, pp. 9-13.

Dewatripont, M. and G. Roland (1995), "The design of reform packages under uncertainty", *The American Economic Review*, Vol. 85/5, pp. 1207-1223, www.jstor.org/stable/2950984.

Dobbie, W. and R.G. Fryer (2011), "Are high quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone", *American Economic Journal*: *Applied Economics*, Vol. 3/3.

Fazekas, M. and T. Burns (2012), "Exploring the complex interaction between governance and knowledge in education", *OECD Education Working Papers*, No. 67, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9flcx2l340-en.

Fryer, R.G. (2011), "Financial incentives and student achievement: Evidence from randomised trials", *The Quaterly Journal of Economics*, Vol. 126/4, pp. 1755-1798, http://dx.doi.org/doi:10.1093/qje/qjr045.

Fullan, M. (2009), "The large-scale reform comes of age", *Journal of Educational Change*, Vol. 1/1, pp. 5-28.

Grossmann, D. (2014), "Secret Google lab 'rewards staff for failure", webpage, BBC News, www.bbc.com/news/technology-25880738 (accessed 27 October 2015).

Hallsworth, M., S. Parker and J. Rutter (2011), *Policy Making in the Real World,* Institute for Government, United Kingdom.

Harlem Children's Zone, (n.d.), "From Harlem Children's Zone", webpage, www.hcz.org/?s=From%20Harlem%20Children's%20Zone (accessed 1 March 2014).

Harrisson, G.W. and J.A. List (2004), "Field experiments", *Journal of Economic Literature,* Vol. 42/4, pp. 1009–1055, http://dx.doi.org/10.1257/0022051043004577.

Heilmann, S. (2008a), "From local experiments to national policy: The origins of China's distinctive policy process", *The China Journal*, Vol. 59, pp. 1-30.

Heilmann, S. (2008b), "Policy experimentation in China's economic rise", *Studies in Comparative International Development,* Vol. 43/1, http://dx.doi.org/10.1007/s12116-007-9014-4.

Hopfenbeck, T. et al. (2013), "Balancing trust and accountability? The assessment for learning programme in Norway: A Governing Complex Education Systems case study", *OECD Education Working Papers*, No. 97, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3txnpqlsnn-en.

Lepi, K. (2013), "How one school uses social media to empower parents and students", Edudemic, webpage, www.edudemic.com/social-media-parents-students/ (Accessed 20 March 2014).

Lipton, D. and J. Sachs (1990), "Creating a market economy in Eastern Europe: The case of Poland", *Brookings Papers on Economic Activity,* Vol. 1, http://dx.doi.org/10.2307/2534526.

Loeb, S. and D.N. Plank, (2008), "Learning what works: Continuous improvement in California's education system", *Policy Brief*, Vol. 8/4, Policy Analysis for California Education, Berkeley and Stanford.

Mason, M. (2008), "What is complexity theory and what are its implications for educational change? ", *Educational Philosophy and Theory***,** Vol 40/ 1, pp. 35-49.

Ministerie van de Vlaamse Gemeenschap [Ministry of Flemish Community] (2009), *Evaluatierapport van de Proeftuinprojecten (2005-2009)* [Assessment report for the experimental projects 2005-2009], Ministry of Flemish Community, Brussels.

Muralidharan, K. and V. Sundararaman (2010), "The impact of diagnostic feedback to teachers on student learning: experimental evidence from India", *Economic Journal*, Vol. 120/546, pp.187-203.

O'Day, J.A. (2002), "Complexity, accountability, and school improvement", *Harvard Educational Review*, Vol. 72/ 3, www.hepg.org/her-home/issues/harvard-educational-review-volume-72-issue-3/herarticle/_48.

OECD (2014), "The weight of nations", *Trends Shaping Education Spotlight 2*, OECD Publishing, Paris, www.oecd.org/edu/ceri/SpotlightBodyandSociety.pdf.

Page, S.E. (2006), "Path dependence: An essay", *Quarterly Journal of Political Science*, Vol. 1/1, pp. 87-115, http://dx.doi.org/10.1561/100.00000006.

Popper, K.R. (1957), *The Poverty of Historicism,* Routledge, London.

Sachs, J.D. (2006), *The End of Poverty: Economic Possibilities for Our Time,* Penguin Group, USA.

Schwerdt, G. et al. (2011), "Effects of adult education vouchers on the labor market: Evidence from a randomised field experiment", IZA, Bonn, www.ideas.repec.org/p/iso/educat/0094.html.

Snyder, S. (2013), "The Simple, the complicated, and the complex: Educational reform through the lens of complexity theory", *OECD Education Working Papers*, No. 96, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3txnpt1lnr-en.

Swedish National Agency for Education [*Skolverket*] (2011), "Municipal responsibility in practice", Swedish National Agency for Education, Stockholm.

Taylor, C. (2009), "Choice competition, and segregation in a United Kingdom education market", *American Journal of Education,* Vol. 115/ 4, pp. 549-568.

Thompson, M. and D. Wiliam (2008), "Tight but loose: A conceptual framework for scaling up school reforms", in C. Wylie (ed.), *Tight but Loose: Scaling Up Teacher Professional Development in Diverse Contexts*, Educational Testing Service (ETS), Princetown, New Jersey.

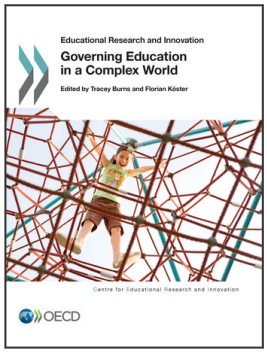
Van der Klaauw, W. (2002), "Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach", *International Economic Review*, Vol 43/4, pp. 1249–1287.

Van Twist, M. et al. (2013), "Coping with very weak primary schools: Towards smart interventions in Dutch education policy", *OECD Education Working Papers*, No. 98, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3txnpnhld7-en.

Van Zanten, A. (2003), "Middle-class parents and social mix in French urban schools: Reproduction and transformation of class relations in education", *International Studies in Sociology of Education*, Vol. 13/ 2, pp. 107-123.

Waslander, S., C. Pater and M.V. Weide (2010), "Markets in education: An analytical review of empirical research on market mechanisms in education", *OECD Education Working Papers*, No. 52, OECD Publishing, Paris, http://dx.doi.org/10.1787/5km4pskmkr27-en.

Williams, R.J. et al. (2002), "Two degrees of separation in complex food webs", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 99/20, pp. 12913-12916, http://dx.doi.org/10.1073/pnas.192448799.



From:

Governing Education in a Complex World

Access the complete publication at:
https://doi.org/10.1787/9789264255364-en

Please cite this chapter as:

Blanchenay, Patrick and Tracey Burns (2016), "Policy experimentation in complex education systems", in Tracey Burns and Florian Köster (eds.), *Governing Education in a Complex World*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264255364-10-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.