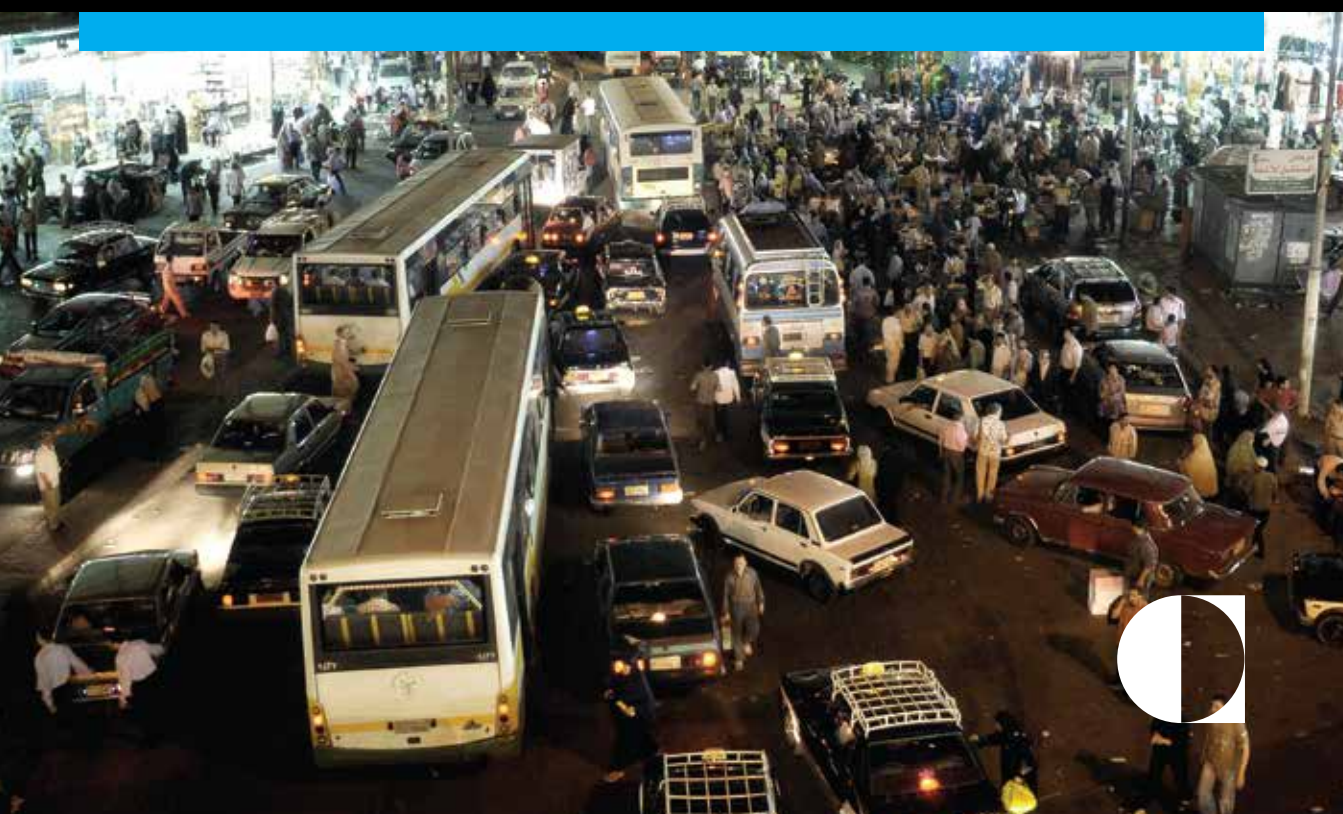
Rachel Kleinfeld

IMPROVING DEVELOPMENT AID DESIGN AND EVALUATION

Plan for Sailboats, Not Trains



IMPROVING DEVELOPMENT AID DESIGN AND EVALUATION

Plan for Sailboats, Not Trains

RACHEL KLEINFELD

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue, NW
Washington, DC 20036
P: +1 202 483 7600
F: +1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org/pubs.

Cover photos: GettyImages.com (top) and ezequiel-scagnetti.com (bottom)

TABLE OF CONTENTS

ABOUT THE AUTHOR

RACHEL KLEINFELD is a senior associate in the Democracy and Rule of Law Program at the Carnegie Endowment for International Peace, where she focuses on issues of security and governance in post-conflict countries, fragile states, and countries transitioning to democracy. As the founder of the Truman National Security Project, she spent nine years as CEO of a movement of national security, political, and military leaders working to improve the security policies of the United States.

Kleinfeld is the author of *Advancing the Rule of Law Abroad: Next Generation Reform* (Carnegie Endowment for International Peace, 2012), which was chosen by *Foreign Affairs* magazine as one of the best foreign policy books of 2012. She has been featured in the *New York Times*, the *Wall Street Journal*, and other national television, radio, and print media.

ACKNOWLEDGMENTS

This report benefited from a host of early readers based at international organizations and foundations, as well as specialists in complexity theory including Henk-Jan Brinkman, Wade Channel, Larry Kramer, Nick Menzies, Gary Milante, Scott Ortman, Steven Teles, Erwin van Even, Christian Voelkel, and my Carnegie Endowment colleagues Tom Carothers, Sarah Chayes, Diane de Gramont, and Nick Wright, as well as Holly Yeager, our terrific copy-editor. I also had the pleasure of presenting the ideas to colleagues at workshops convened by the World Justice Project and the University of Chicago Law School; the Clingendael Foundation and the Netherlands' Permanent Mission to the United Nations; and the Justice Assistance Network of the UK government. While all faults are my own, I am grateful for the generosity of so many colleagues in offering insightful comments, searching questions, disagreements, and improvements.

The Carnegie Endowment's Democracy and Rule of Law Program is grateful to the UK Department for International Development for its research support that helped make the writing and publication of this report possible.

SUMMARY

THE DEVELOPMENT FIELD increasingly looks to sophisticated metrics to measure impact. Simultaneously, practitioners are recognizing that most development programs must engage with politics and policy. Unfortunately, the measurement techniques gaining popularity are those least able to determine how to implement political reforms. Effective reform efforts require planning for and measuring change that is nonlinear and nonincremental. Complexity, or systems, theory offers insights for improving program design and evaluation.

THE NATURE OF POLITICAL REFORMS

- In more political development programs, opponents may contest both ends and means. Programs that get adopted are rarely technical best practices, but rather those that amass the most political support.
- The presence of opposition actors means that reforms are frequently followed by counterreforms. Change swings back and forth. Measuring success at only one point in time means little for whether a reform will be sustained.

- Political variables are interdependent, but popular measurement tools such as regressions and randomized controlled trials assume variables can be separated. These techniques can determine which interventions are most effective—but not how to get those programs implemented.
- Designing programs that alter the underlying rules of political and social systems is the key to successful reform.

RECOMMENDATIONS FOR DESIGNING AND MEASURING REFORM EFFORTS

Design programs and funding to anticipate counterreforms and multiple battles. Opposition learns, too: techniques that worked at one point may fail at another.

Engage local partners who can amass broad coalitions. Avoid making groups overly beholden to donor agendas that can cost them local support. Measure programs based on whether they have created long-term, broad coalitions and/or elite influencers with real political power who are growing stronger.

Ensure flexibility for programs and budgeting, and expect changes. Test hypotheses throughout a program's life cycle. Design contracts to enable closing projects and moving funds among projects so that acting on what works does not carry a stigma or lead to perverse incentives.

Prepare for windows of opportunity before they open. Invest in coalitions, policy development, and social networks ahead of time.

Determine whether programs have shaped the rules of the system to make change easier. Programs that enable organizing, increase transparency and public voice in policy, reduce violence against reform advocates, and increase avenues to power are types of systemic changes that allow reform.

Measure reform based on the space of the possible. Look at all the potential options in a policy space, including possible counterreforms, not just the currently ascendant policy.

INTRODUCTION

While out hiking with a friend recently, I found myself describing the difficulties of improving the rule of law in Honduras, an issue I was wrestling with for a donor assessment. After miles spent listening to me rattle off problems of corruption, police brutality, entrenched elites, gangs, narcotraffickers, and assassins for hire, my friend asked, "Don't you get discouraged working on these intractable problems?"

I stopped midstep. The development community exists to overcome big problems of poverty and governance. These problems are hard—I wasn't expecting much positive change in Honduras anytime soon—but they are not necessarily intractable. Changes that make countries wealthier, more inclusive, and better governed happen regularly.[1] Countries from Portugal to South Korea leapt from poverty to wealth in a generation. Multiple U.S. municipalities pulled themselves out of entrenched, institutionalized corruption from the 1880s to the 1930s. A social movement ended slavery in Britain and, eventually, most of the world.

It's easy to see these successes in hindsight. But in the thick of the fight, things are not so clear. Progress is often two steps forward then two or three steps back—and sometimes it moves sideways. Forward momentum is not inevitable; reforms can be stymied by opposing viewpoints, and countries can simply get stuck, unable to resolve important challenges. A sudden event may harness a public mood and create an opportunity, or block one.

In other words, progress looks less like a freight train barreling down a track, whose forward motion can be measured at regular increments, and more like a sailboat, sometimes catching a burst of wind and surging forward, sometimes becalmed, and often having to move in counterintuitive directions to get to its destination.

Today, the vast majority of development projects require engaging in the realms of policy, power, and politics. And, whether you are funding change, fomenting it, or opposing it, the nonlinear nature of this kind of reform can make it very hard to know whether you are on the right track, and how to measure whether you are achieving your goals.

As development practitioners and donors begin to acknowledge the importance of politics to their work, they are grappling with how to design and evaluate their programs. Metrics, indicators, and assessment tools have proliferated in the last two decades, driving dollars and strategies.[2] Yet many people engaged in social and political reform know that there is something wrong with how the development community is designing and evaluating its efforts. Relying on the wrong type of metrics risks moving the field toward programs that are less suited to addressing political reforms.

For instance, if you worked at a foundation that was providing funding to gay rights groups in the United States, the time chosen to measure a program would have made a vast difference in whether you saw your efforts as succeeding or failing. In 2006, things would have looked pretty bleak. Same-sex marriage was illegal in every state of the union except for Massachusetts, and over the previous two years, 24 states had passed constitutional amendments to ban gay marriages or civil unions. How would you know whether the huge political losses would galvanize change, or cause despair and ennui in the movement? Perhaps you had been at this work since the Stonewall riots of 1969, when the gay rights movement really began. Should you change your strategy—or keep hammering away? In 2006, how could you determine what tactics to use and how to evaluate those programs, so that in 2014, forty-five years after your activism had started, you could take some credit for the United States moving from homosexuality itself being illegal in many states to some form of same-sex marriage being legalized in more than half of them?

Not only is the trajectory of political and social change murky in the midst of the struggle, it is also not clear when the struggle ends. If you were on the other side of the gay rights movement, 2006 would have appeared to be a moment of celebration. Yet eight years later, counterreforms had overturned all that had been achieved by those working to protect traditional marriage. If funders or reformers were measuring success at that apex, how could they have guessed at the backsliding that was about to occur? And the struggle is not over: no one knows what the next chapter holds.

Forty-five years is not an uncommon time span for development efforts that are politically fraught. The World Bank's 2011 *World Development Report* suggests that the fastest rule of law change takes forty-one years.[3] Studies of deep political and social change

by Douglass C. North, John Joseph Wallis, and Barry R. Weingast put the timeline at around fifty years.[4] Lant Pritchett, Michael Woolcock, and Matt Andrews come up with even slower rates of change.[5] How do development agencies, foundations, and activists design programs that can keep up momentum and funding over this length of time? And how can they measure their potential success in a realistic timetable, particularly when they know that change doesn't move in a smooth, straight line?

Designing effective programs that involve politics and evaluating these processes appropriately matters, because development practitioners are realizing that politics is involved in ever-broader areas of development work. As Thomas Carothers and Diane de Gramont have written, "The overdue recognition that development in all sectors is an inherently political process is driving international aid providers to try to learn how to think and act politically."[6] Daron Acemoglu and James Robinson have written numerous articles and a best-selling book describing how most economic reform requires political reform if it is to take hold.

The World Bank, like other donors, is struggling, caught between Articles of Agreement that mandate that "the Bank and its officers shall not interfere in the political affairs of any member" and the recognition that its development goals now include "protection of global public goods, governance, and institutions, as well as issues such as inclusion and cohesion, participation, accountability, and equity"—among the most politically contentious issues in any society.[7] Its attempts to square this circle reveal an organization trying to address what are now known to be underlying causes of poverty while tying programs into knots to do so legally.

Other development organizations are caught in similar conundrums. Nongovernmental organizations (NGOs) that find themselves excoriated by the world's autocrats often claim for their own safety that empowering women and minorities, abetting civil society, or providing services to the poor are not political activities. The autocrats they might unseat with these programs that enhance equity and transparency think otherwise.

The reality is that most development involves politics in some way. Sometimes development projects engage in bureaucratic or small-scale politics, such as deciding where to place a village well—near the chief's home, where he has kindly donated land but could then control use, or near the poorer people of the village. This is not the level of politics under discussion.

Instead, this report is concerned with larger-scale political and policy engagement. It applies to the subset of the development world that is engaged in democratization, a community that it has long been clear is involved in politics. But it also has a broader ambit. The development community has more recently become involved in governance, anticorruption, transparency, and rule of law programs. These efforts universally affect laws and policies, and nearly all face opposition—and thus all are political.

Finally, many, perhaps most, large-scale socioeconomic development programs also require political engagement. Politics is the process of making decisions about the rules that govern a society and the use of public resources. These decisions are never purely technical. Even if the end goal is not to affect a regime or a political party—but simply to build a road, help girls get education, or reduce child mortality—interventions that affect how public resources are produced, who gets those resources, who makes allocation decisions, and what rules govern relations between those who make decisions and those who don't are all political interventions.

This report concerns program design and evaluation for all three types of engagement: in other words, most of development work.

As the fight over the 2015 Millennium Development Goals illustrates, the decisions that funding agencies make about what to measure can determine their activities and their program designs. It's important to measure the right things in order to incentivize programming that works.

Yet people who work on social and political reform acknowledge that the development community's current approach is inadequate. Matt Andrews has been leading a charge against so-called best practices and rigid program design and for what he and others at Harvard's Kennedy School call a Problem-Driven Iterative Adaptation (PDIA).[8] Duncan Green's Oxfam blog ran a "wonkwar" debate on various evidence-based approaches to measuring political change.[9] The Developmental Leadership Program, an international research initiative based at the University of Birmingham in the UK, has been holding conferences and writing white papers on better ways to measure the politics of development.[10] The *Stanford Social Innovation Review* ran a series on the design and measurement failures of strategic philanthropy—kicked off by none other than the leaders of the movement that advocated the reigning "logic frame" structure of strategic philanthropy in the first place.[11]

Interventions that affect how public resources are produced, who gets those resources, who makes allocation decisions, and what rules govern relations between those who make decisions and those who don't are all political interventions.

Meanwhile, in the field, programs are faltering because they give lip service to politics, but are not matching their words with altered operations and metrics. In West Africa, the director of the consortium that managed a series of major donor-funded governance reform

programs told me: "None of these programs are going to work. All this money, all these political economy analyses—none of it will make change. Because [the Western donors] are scared to do what would really matter. They fund NGOs with small bases of support in the capital, but won't fund the mass groups that are out there to fight politically."

Furthermore, while there is near-universal rhetoric about the need for flexibility, it is rare in practice. In Chile, the catalysts for what was arguably the most successful judicial transformation in South America described how they had to fight with their donors just for the freedom to take money they had already been granted for legal reform and put it to a slightly new end when the politics on the ground changed.

Many practitioners and scholars, in other words, know that politics increasingly matters to their work. Many development organizations have accepted that their work is political. But too often, the development community's procedures for designing and measuring their programs are not suited to political reform. This report seeks to close this growing gap by addressing the following questions:

1. How do reforms that require political engagement differ from traditional technical reforms?

2. Why is political engagement different, and what are the implications for design and evaluation?

3. How should development programs that engage politics be designed?

4. How can those who fund or implement such programs evaluate whether their efforts are contributing to reform?

The growing chasm between what is understood and what is actually enacted in the majority of programs has also bifurcated the development community. If you are a development practitioner who already thinks that politics is paramount to developmental reform, change is nonlinear, the main obstacle to implementation is not what to do, but how to get it enacted against opposition, and that best practices therefore tend to fail—and your institution is using that understanding to investigate alternative methods and create better program designs and evaluations, feel free to skip to the third and fourth sections of this report.[12]

If, however, you are a practitioner stuck between your own appreciation of the political nature of development and your institution's continued use of a technocratic set of standard operating procedures, the next two sections can provide a language for talking about why fundamental change is needed. And if you feel that the whole political turn in the development field is incorrect, or you are comfortable with today's measurement techniques, then I particularly welcome your engagement on all parts of this report.

WHAT MAKES POLITICAL REFORM DIFFERENT

The paradigm of the development field since the 1950s has been one of delivering services, either directly, as with humanitarian aid, or by assisting local governments with infrastructure and technical support. The goal is clear and assumed to be universally good (providing more food, better healthcare, improved transportation). The question is how to do it in the most efficient manner. That technical mind-set from the worlds of engineering and economics has informed how projects are designed and measured even as it has become clear that many of these efforts are political as well as technical. As Steven Teles and Mark Schmitt write about this conundrum in the United States:

> Foundations, universities, and government have developed sophisticated tools for evaluating service-delivery programs and smaller-scale tests. These methods range from controlled experiments, to the identification of best practices that seem to be transferable from one successful program to another, to a more malleable form of evaluation based on assessing the "theory of change" underlying an initiative. The development and implementation of these tools, often on a large scale, constitutes a growing industry of its own.... These sophisticated tools are, we will show, almost wholly unhelpful in evaluating efforts at advocacy.[13]

Development that engages politics is different from traditional service-delivery programming in three important ways:

1. The end goals or methods are in dispute, meaning best practices are unlikely to emerge.
2. There is an opposition fighting back, resulting in reform that is nonlinear, non-incremental, and difficult to measure along a straight line.
3. The variables are interdependent, making measurement techniques that require holding other things constant ineffective.

THE END GOALS OR METHODS ARE IN DISPUTE

In any kind of development work, a typical evaluation to determine the best way to do something starts with researchers who know the goal and are looking for the most efficient way to get there. They ask questions such as "What policing method is most effective at reducing crime?" or "How can we increase the number of patients taking their medicine?" These are useful questions to answer once a policy has already been adopted and has political will behind it.

But in the political realm, before a policy has been agreed upon, the end goals themselves are often in dispute. This can easily be forgotten when programs and sophisticated evaluations are designed—after all, who is against more girls reading, more parts of a country being electrified, or more kids living past the age of five?

Sometimes people are against such goals for venal personal reasons. Leaders of what are known as extractive, neopatrimonial, or limited access economic-political orders are notorious for refusing to develop parts of their countries because doing so would threaten their hold on power or their personal gains from what are supposed to be public goods. Development agencies may want to improve transportation links to a poor part of a country to help that region get goods to market, for example. However, as occurred in Guatemala in the mid-1800s and persisted through modern times, the country's cartel-like main business organization may collude with the political leadership to block a new road that would empower new businesses. These elites want to continue to reap monopoly profits from the existing businesses, ports, and transport links they run, not to enable competition.[14]

At other times, it's a question of priorities. No one disputes that the end goal is good and the methods are sound—but a multitude of desired ends are competing for time and budget, and the reform goal preferred by development agencies is not the one backed by the loudest and most powerful voices in the country. For example, in a controlled study, social scientists can determine whether class sizes, skilled teachers, or better curricula are optimal for increasing reading scores. But when it comes time to get that method

accepted as government policy, those findings may run into passionate advocates who don't want to lose arts education, teachers' unions that resist merit-based firing, science and math aficionados who have nothing against reading but want meager school budgets and time spent on their topics, and so on. The failure of the school reform effort that began in 2010 in Newark, New Jersey, despite ample funds, a supportive political establishment, and celebrity proponents, is a case in point.[15] In situations like this, coalition politics often cobble together agreements to push for changes that are suboptimal from a technical standpoint, but that can garner broad agreement or are better for a particularly powerful group. Such political processes frequently force reform programs to move sideways rather than strictly forward or backward.

And sometimes there may be broad agreement that the end is good and a top priority—but so much disagreement on the methods to achieve it that change stalls and no method at all gets adopted. For instance, problem-oriented, hot-spot policing is a best practice that has emerged from a series of statistically strong studies in the United States. It is now well established that most crimes are committed by a very few people, in a very few places, at specific times. Put more police on the streets in those areas, have them work with the community to solve local problems, and the best practice suggests that crime will decrease. These findings are strong, repeated, and have probably contributed to reduced crime in many municipalities.

Yet getting best practices adopted is difficult, even in places that accept that reforming policing is a top priority, because the methods themselves may meet with opposition. For example, those who try to take problem-oriented policing into new municipalities—whether in a U.S. city struggling with policing issues such as Ferguson, Missouri, or in a country such as South Africa—are likely to run into lobbying groups who complain that concentrating police in areas with high crime leads to racial profiling. Pro-gun groups will claim that limits on firearms reduce individual rights and prevent self-protection. Advocates for the poor will claim that funds for law enforcement and the prison system should be allocated to social interventions. The research itself will be lost or manipulated in the melee. In some cases, the very goal of reducing homicides will be challenged as people question the social justice of widespread arrests of one group, call it an unacceptable price to pay for lowering crime, and complain of a growing police state. This is not just the typical problem of transplanting reform. It is that agreement on end goals does not imply any agreement on the methods of reform, and many ideological disagreements are around methods, not ends.

Getting best practices adopted is difficult because the methods themselves may meet with opposition.

Insisting on a best practice may make any reform impossible to attain, as the technocratic ideal is only occasionally in anyone's political interests. Instead, when ends or methods are in dispute, a realistic policy will only emerge from the process of contestation. That is because most political reforms happen when common ground is found among those with somewhat different views and a coalition is cobbled together that is big enough to get a policy passed, even if it pleases no one perfectly. While best-practice reforms are rarely adopted as is, savvy reformers expect a second round, and they know that they will revisit the original dispute through secondary fights, with new alignments of the original forces in the debate and others they can mobilize. If a country is under pressure from international donors, best-practice reforms may get adopted into law or policy—but they are then usually ignored, because they have not been through this process of political accommodation.

Because best practices are already settled, based on technical knowledge often found outside the country or city in question, they can't accommodate this process of compromise between competing factions with differing views. In fact, even when a best practice works at one time period or in one part of a country, it may not find fertile ground at a later time or in another part of a country—in part because the opposition may learn from the first wave of reform and push back before it can spread. Measuring reform based on how closely it approximates best practices sets reformers up for failure.

Measuring reform based on how closely it approximates best practices sets reformers up for failure.

THERE IS AN OPPOSITION FIGHTING BACK

Because there are nearly always people who disagree with the goal of a political reform or the means selected, there is usually an opposition working actively against reformers. There are also intermediate forces: bureaucrats, citizens, unions, businesses, media interests, individual opinionmakers—each with its own interests that can be organized for or against a reform. To quote a high-level police officer charged with fighting corruption in Nigeria: "If you fight corruption, it fights you back."[16]

The dynamic of multiple sides to a fight, each gathering its forces and applying pressure, means that changes that involve political decisionmaking do not move in a straight line. Reforms tend to be met with counterreforms, and movement tends to swing back and forth.

The Pushback Against Vaccines

The movement to vaccinate children in much of the world made huge strides from the 1950s through the 1990s. Some countries required these vaccinations by law, while others just made them standard practice. Smallpox was eradicated, and it looked like polio might be next. In the West, many other once-common childhood diseases simply no longer occurred. But in 1998, a British doctor published an erroneous study linking the measles, mumps, and rubella vaccine to autism. A backlash against vaccination started in the developed world as parents pushed policymakers and doctors to allow them to opt out of the vaccines. Now the West faces a surge of once-prevented diseases, such as measles and whooping cough.

Meanwhile, proponents of political Islam were rising in much of the developing world. They claimed that vaccines were Western imports intended to hurt Muslim children. Their argument gained additional currency when the Central Intelligence Agency used a fake vaccination drive to help identify Osama bin Laden. A backlash against healthcare workers in places like northern Nigeria and Pakistan has increased the presence of polio and other diseases that had nearly been eradicated.[17]

This pendulum pattern of reform and counterreform characterizes all development efforts that affect policy and politics. And it wreaks havoc with traditional program design and measurement. For instance, the most sophisticated and widespread approach to program design today, the logic frame, calls for a determining a theory of change and then a set of activities that backs that theory, along with intermediate goals, and regular, measurable steps toward progress along the way that are defined at the beginning of a program.

But that is just not how reform occurs. Instead, nothing may happen for a long while as forces amass their power—and then sudden change may take place when a window of opportunity arises. Or linear, incremental reforms can suddenly face rollback when the opposition gains a moment of opportunity. It makes sense to think logically about a theory of change and how to get there in order to shape initial activities. But rather than expecting programs to remain static over time and serve as guides for evaluation, one must expect that, like battle plans, they won't survive contact with opponents.

The reform-counterreform dynamic of development efforts that involve politics means that programs must be designed to take into account the need for ongoing waves of a fight. This is particularly true of funding. If a program declares success and ends after the initial reform and its funding dries up, reformers are left utterly unprepared to fight the next, almost inevitable, battle.

The Danger of Ignoring the Next Fight

In Georgia after the Rose Revolution of 2003, donors slashed funding for NGOs that had agitated for accountable, noncorrupt government, and instead moved to help the new government enact a series of governance reforms. The reforms succeeded in reducing petty corruption, improving customs and traffic policing, and bringing change in other important areas. But the government grew overconfident in its own ideas, and autocratic in its implementation. When the government started to control the media, judiciary, and other oversight bodies, and to use taxation authorities to crack down on legitimate businesses, NGOs were weak and unable to make their voices heard to ensure accountability over the long term. As a result, Georgia built a more functional state—but did not increase the rule of law.

Likewise, in Chile, landmark criminal justice reforms in the late 1990s and early 2000s reduced the percentage of prisoners stuck in pretrial detention before they had been found guilty of any crime. A few years later, the media highlighted prisoners released before trial who had committed new crimes, and politicians excoriated the judiciary as being too liberal and overly concerned with the rights of perpetrators, not victims. Yet the array of institutions that had rallied public support for the initial reforms now lacked the cohesion and funding to fight the backlash of legislation that followed.

Most development projects, if measured at all, are evaluated within about a year of their closing date. But for political and social change, policy durability matters, and that often can't be determined in just a year.[18] Because counterreforms are so frequent, any given moment of reform is often, metaphorically, the end of a battle, not the end of the war. This means that for political reforms, timelines for measurement must be lengthened. Moreover, no measurement at any single point in time can present this dynamic picture. Instead, multiple measurements at different times will be most useful, and intermediate variables must be found that are not simply linear projections back from a settled goal.[19]

Most evaluation systems are set to measure the equivalent of a train progressing down a track: a straight line that starts a little slowly and then gains speed, with clear checkpoints along the path that should be hit at specified times. Social and political reform looks like a sailboat tacking toward its destination, sometimes over the course of fifty years. Like Odysseus's famous journey home, it entails odd bedfellows, unexpected diversions, eddies of inaction, and moments of opportunity to surge forward.

THERE IS NO INDEPENDENT VARIABLE

Increasingly, development projects are looking to sophisticated regressions and randomized controlled trials (RCTs) as the most favored forms of measurement. Both are based on methods of analysis that must alter one variable while holding others constant, either by aggregating enough people that individual differences are no longer statistically relevant, or by using specialized mathematical techniques such as dummy variables.[20] Yet in political change, variables are interdependent. It is impossible to change just one thing. While a regression or RCT trial can be run for a pilot, the findings often fall apart in the real world.

Indeed, the robust testing designs prized by many aid agencies are useful for measuring what would work best in a technocratic world absent politics, or where a policy has already been determined. Yet according to Robert Jervis, former president of the American Political Science Association, they are the wrong form of measurement for answering the "how" questions that aid agencies need answered to get policies adopted and implemented.[21]

In both regression-based and RCT diagnostics, no matter how sophisticated the techniques employed, independent variables can be separated from all other variables and tested in isolation, with the assumption of a Newtonian universe in which, like a watch, pieces can be disassembled and studied separately. Those assumptions are true for all sorts of complicated phenomena, such as how to build a dam or power plant. But these assumptions fall apart when variables become interdependent.

Consider the case of outside funders trying to figure out how to reduce violence in Colombia. In the 1980s, Colombian drug cartels began to buy ranches to convert drug proceeds into real estate, often in areas where left-wing guerrillas were fighting and had driven down prices. The narcotraffickers allied with local landowners to form paramilitary self-defense groups to protect themselves and their businesses from kidnapping and extortion, and they often murdered guerrillas and anyone believed to be sympathizing with them. However, while the drug cartels were fighting the guerrillas, they were also cooperating with them in the business of growing and transporting cocaine.

The paramilitaries also worked closely with Colombia's police and military, attracting former law enforcement officers to their ranks and using government-provided intelligence to fight the guerrillas.

At the same time, some left-wing guerrillas had begun to operate in the poor suburbs of major cities, sometimes wiping out gangs and petty criminals to ingratiate themselves with the local population. Many in the local police appreciated this form of crime control, and while officially fighting violent left-wing militias, the police often cooperated with them or turned a blind eye toward their behavior. Drug cartels' paramilitaries copied the strategy, with the result that the Colombian government was often both

working with and fighting against the drug cartels, and with and against the guerrilla groups at the same time, while these groups warred against each other over turf and illicit business opportunities.

In the world of development politics, simply finding accurate data on factors such as who is committing violence is hard enough. Isolating independent and dependent variables, such as whether a particular program reduces paramilitary violence, is nearly impossible, because the variables are interdependent. Meanwhile, each action yields side effects that would be lost in methodologies that fail to look for them—such as the collusion and corruption between police and criminals that occur when law enforcement officers turn a blind eye to murders of unwanted members of society.

Randomized controlled trials and regressions both have useful contributions to make in determining more technical parts of the solutions to these kinds of puzzles. Their findings can help shed light on the best ways to run a disarmament, demobilization, and reintegration (DDR) program to reduce paramilitary recidivism, or the best program to curb the growth of gangs.[22] But they cannot offer much of use in determining how to get a polarized political scene to accept the best DDR program, or how to stop the Colombian government from collaborating with drug cartels against guerrillas. These methodologies are not suited to finding answers to these political process questions, the "how" as opposed to "what" questions.

Using the wrong type of methodology to answer necessary questions is not useful, even if the studies themselves are rigorous and well done. The development community needs to consider other paths to evaluate and design programs.

USING COMPLEXITY THEORY TO UNDERSTAND AND MANAGE REFORM

So far, I've enumerated how development that involves politics is different from technical programming. It's time to start looking at why it's different, and what that means for creating and measuring change.

Politics—and the development work that must engage with politics—is complex. That's not an admission of defeat, but a technical term that describes a very distinct kind of problem that has been the focus of a branch of scientific inquiry across fields such as physics and biology for the past thirty years. This interdisciplinary scientific method, known as complexity theory, has already been brought into the social sciences, where it is known as systems thinking.[23] It has revolutionized fields such as urban design, transportation, and epidemiology. Now some aid agencies are trying to incorporate systems thinking into their work. The payoff for politically informed development could be vast.

WHY POLITICAL CHANGE IS NONLINEAR AND HARD TO MEASURE

Complexity theory says that physical and biological systems can be divided into three categories. Ordered systems are predictable and linear: causality moves in a single direction. These systems are also Newtonian: each part can be separated from the whole and

examined, then aggregated again to understand how the whole works, like a Lego building. Any measurement regime that requires independent variables, or yields findings that do not depend on a particular time and context, is predictive only in such linear, Newtonian systems.

Another type of system is chaotic. Chaotic systems are subject to constant, mathematically unpredictable change. Human systems devolve into chaos perhaps in the midst of an active battlefield during war, when the rules of the game are constantly switching and randomness plays a significant role in determining who survives. But it happens only rarely.

Most of the time, social and political life happens in the arena in between: complex systems. A complex system is defined by having many autonomous actors that have multiple interactions with each other. The actors are interdependent: the actions of one influence the other, which in turn, influences the first. And their actions together can influence the whole system, so that the agents shape their environment. This means it's not Newtonian: such a system can't be disaggregated—taking the pieces apart to study each one loses the interaction effects among them. In the world of development, these interaction effects between the pieces are the "formal and informal rules and institutions" that are crucial to understanding systems, which North and the New Institutional Economists have been speaking of for the last two decades.

Many systems are complicated, but not complex. For instance, building a car or a nuclear reactor is not simple, but no matter how many parts are involved, each variable is Newtonian: they all fit together and can be taken apart, and they don't somehow change by being placed near one another. Building a healthcare system, an educational system, or a democracy is different: changes in one factor affect many others, which in turn affect the first.

Scientists have determined certain mathematical facts about how complex systems work in the physical world. Understanding these tendencies has changed how a variety of other social fields operate, and has greatly increased effectiveness in tracking diseases, reducing traffic congestion, and improving other complex systems. This knowledge can also provide insight into how to design and measure programs that must affect the political world.

Feedback Loops

One reason that complex systems don't change in a linear way is because they have self-reinforcing and self-defeating feedback loops. With feedback loops, change can start small and appear linear at first. Yet it starts to snowball as the loop feeds back into itself. Depending on what is happening, these self-perpetuating feedback loops can create virtuous—or vicious—cycles that can cause reform to take off or make it move backward.

In Colombia, rural violence metastasized during the bloodletting known as *La Violencia*. After the assassination of a popular presidential candidate in 1948, liberals formed guerrilla bands and conservatives formed self-defense forces that engaged in years of violence.

As the political killings spiraled out of control, others turned to violence to settle personal disputes, knowing that they would have impunity given how overwhelmed local authorities and the judicial system were. In fact, the more crimes, the more impunity each criminal had. Landowners began to kill and threaten peasants in order to drive them off the land; domestic violence escalated; criminals killed competitors; businessmen contracted to kill rivals. Over time, not engaging in violence to get what one wanted seemed to be more unusual than using force.

The feedback loop did not stop there. Many of those who would later lead drug cartels grew up in this atmosphere of brutality. When the cartels began to form in the 1970s, they recruited many of their assassins and thugs from the pool of young people in city slums, some of whose families had been dispossessed from their rural homes during *La Violencia* and forced to move to the cities. The cartels could also hide business-related violence against rivals under the more popular guise of the paramilitary self-defense groups that had a long tradition of support in Colombia.

Tipping Points

At this point, complex systems exhibit another typical form of physical behavior. They have tipping points, or phase changes, such as when water freezes into ice or heats into steam. Individual behavior starts to align within a greater force created by the many choices of other individuals. It becomes harder to fight that social force, and easier to go along with the herd, until there is an entirely new structure. In other words, each person's behavior is not formed in a vacuum: as more and more people act a certain way, it becomes harder to act differently, leading more people to act in that way, until the system tips, creating rapid and unpredictable changes.

This phenomenon was evident in Colombia when the military and police started to collaborate with paramilitaries that were sponsored by narcotraffickers. These paramilitaries used dirty tactics against guerrillas that were illegal for agents of the state. Some military and law enforcement members retired and joined the paramilitaries, others moonlighted with the paramilitaries while on active duty, and others passed on intelligence and weaponry to the groups.

Those in the military and police who were unwilling to engage in the paramilitaries' violence themselves or to ignore the violence of their peers found themselves in a dangerous situation. The system had tipped so that many members of their organizations were now

acting outside the law. If a soldier spoke out against collaborating with paramilitaries, he was likely to be killed by his own side, and have it blamed on a shoot-out with guerrillas. Instead of generally honest forces with a few bad apples, these organizations were corrupted and couldn't be trusted to self-police. Meanwhile, the many honest members of the military and police were scared into a code of silence.[24]

Feedback loops and tipping points are part of why measuring incremental reform doesn't work in these systems—change for better or worse can be slow for a while, and then suddenly fall off a cliff. But they also explain why the timing of any measurement is incredibly important in complex systems. Measure just before a tipping point, change can look like very little; measure just after, and it can appear transformative.

Tipping points also mean that the largest events—such as government collapse—often have no particularly exceptional causes. As Ben Ramalingam and his co-authors write, "Every avalanche, large or small, is caused by falling grains which makes the pile just slightly too steep at one point."[25] Big changes can accrete from small, incremental actions, feeding back in on themselves again and again.

This has many implications for program design, including the need to take into account the systems within which people work. For instance, picking a champion to tour a Western country and learn lessons for reform that she will take back to her government rarely yields systemic change by itself. Such reformers drop back into systems that are resistant to change. Simultaneous efforts to help multiple reformers alter the system are as necessary as altering individuals' sense of possibility.

Path Dependence

Because of feedback loops, complex systems tend to be path dependent. Once they start down a path, a feedback loop is created, making it hard to alter those patterns. Each step sets the stage for the next. This means that program design and measurement must take timing into account. Context is not just geographic and cultural, but also time dependent: something that works in one country at one time may not have worked in the past—and may not work in the future, after the opposition develops a strategy and begins to organize to fight back.

In Colombia, former president Álvaro Uribe's expansion of the military and tough military tactics are widely credited with reducing violence during his term in office, from 2002 to 2008. Yet Uribe's election on a law-and-order platform may not have been possible had his predecessor, Andrés Pastrana, not been widely supported for his own platform of peacemaking with the Revolutionary Armed Forces of Colombia (FARC) guerrillas. It was the failure of that peace process in the late 1990s—which the Colombian public saw

as a genuine effort on the part of the government that was rejected by guerrillas—that made voters turn to Uribe's more militarized solutions in the following election. Uribe's strategy also required greater investment in the military, which had long been kept weak and poor to avoid coups. Only after elites had lived through the violence of drug kingpin Pablo Escobar's war on the state, and guerrillas had taken their fight to the cities, were elites willing to support a wealth tax and a larger military budget, instead of simply turning to private protection.

Thus, even if the approach taken by Uribe turned out to be a best practice for reducing insurgency that worked in every case where it was tried, such programming could not be imported into a country where the public was weary of war, or overly distrustful of its security services: the people would not back the policies. Nor would it be fair to declare that organizations pushing for more militarized policies in the Colombia of the late 1990s were failing, even though they had no impact for years while the government negotiated with the guerrillas for peace. The public simply was not interested in more war during those years, and didn't trust the government to wage it well.

Reform initiatives that fail to find a window of opportunity are like seeds planted in the middle of winter. Failure to grow does not mean anything is wrong with the seeds or the gardener: the plant simply can't lay roots until the soil is ready, the rains have fallen, and the sun is out. Timing and path dependence matter to impact. Program design must allow for paths to unfold. Measurement that does not take such path dependence into account risks penalizing effective organizations and rewarding lucky ones.

Measurement that does not take path dependence into account risks penalizing effective organizations and rewarding lucky ones.

Moreover, programs that expect an outcome similar to one achieved by a similar program enacted earlier, even in the same country, may find themselves disappointed. A strategy for change that worked well in the past may fail to achieve the same results once the opposition learns how to fight it and organizes against it, or the public becomes inured to a particular method of engagement. The first ALS Ice Bucket Challenge took the world of charity development by storm in the summer of 2014, as people dumped ice over their heads and, with the help of social media, raised millions to combat a rare disease. But the twelfth such social media fundraising effort may fail to generate much revenue as people grow tired of the technique. Forcing an organization to use the same process that worked in the past may be setting it up for defeat in the future.

The Butterfly Effect

Because feedback loops are always present in complex systems, these systems can be very sensitive to small changes that set such a loop in motion. In popular science, this is known as the butterfly effect—the idea that the flapping of a butterfly's wings in one part of the world can lead to a storm a continent away as that initial flap builds on itself across airwaves.

The butterfly effect means that complex systems are sensitive to small changes, and it also illustrates why idiosyncratic variables matter so much. Variables interact, feed back into themselves, and build up, meaning small differences can magnify and alter historical trajectories. Together with other factors, the butterfly effect explains why windows of opportunity are so important to political change, and why a good program can't simply be enacted by an effective organization whenever it chooses, along a preset timeline.

Vast quantities of development literature cite the need to be open to windows of opportunity or critical junctures, when a particular alignment of actors makes change more possible. But rarely do donors enable their funding cycles and required impact reports to actually abet this needed flexibility. Program implementers are told that they should be flexible and await a moment of opportunity—but they know they must spend down their grants and report progress at regular intervals or risk losing the confidence and support of their funders, meaning that they may not have funds on hand when those moments arise.

The beliefs and personalities of leaders, and the relationships between reformers, can be the difference between reform happening and not occurring. A host of other idiosyncratic variables, including pure luck, can end an otherwise well-designed reform, or assist a poorly designed effort. What this means for evaluation is a big problem: impact tells one little about whether the program itself was lucky or well designed. Measuring based on end-state impact—the goal of most logic-frame measurements—can actually backfire. Excellent organizations and well-designed programs may founder due to unfortunate timing or an idiosyncratic spoiler, while weaker efforts may happen to catch a wave, find themselves on the edge of a tipping point, and get credit for pushing a reform forward.

In Colombia, the demobilization of paramilitary forces has been a major factor in reduced violence since 2002, despite the reconstitution of some of these forces into criminal gangs. The process relied on a host of idiosyncratic variables: The U.S. decision to officially designate the AUC paramilitary coalition as terrorists happened to occur on September 10, 2001, a day before al-Qaeda would move Latin America's drug and violence issues into a far less salient position in U.S. policymaking. The classification led the paramilitaries to seek ways to avoid landing in U.S. prisons. Another idiosyncratic

but essential variable was the fact that Uribe happened to be in power, and the paramilitaries considered his administration to be sympathetic. They strengthened that sympathy through efforts to corrupt leading politicians and secure easy terms of surrender.[26]

Had the U.S. push for extradition not coincided with Uribe's time in office, the paramilitaries may have tried to fight their way out of their predicament. Indeed, the U.S. effort to extradite drug lords in the 1980s and early 1990s had just this violence-escalating effect, as Pablo Escobar and other "extraditables" declared a war on the state. But under Uribe, those accused of similar or worse crimes chose to demobilize. Though many returned to a life of crime in subsequent gangs, these no longer had the power or ideological weight of the paramilitary movement. Did this mean that the tactic of extradition requests was better implemented the second time than the first, or that reformers in the Uribe administration were more effective than those in earlier administrations? Not necessarily: timing, path dependence, and the coincidence of how each set of criminals assessed its prospects vis-à-vis state leaders played an equally important role.

If the goal of measurement is to determine whether to continue funding an effort, or whether a strategy is likely to work in other contexts, impact alone is not the most useful variable. Measurement regimes should also be aware of the need to look for idiosyncratic variables that may have outsized political effects. It's impossible to anticipate all such small and idiosyncratic variables. But knowing that they are so important to outcomes is key to judging the success of a program and separating what was good or bad luck from good or bad design.

KEYS TO CREATING AND MEASURING CHANGE IN COMPLEX SYSTEMS

Everything about complex systems seems to make any process of design and measurement impossible. Change is nonlinear; it moves in nonincremental ways and tips and metastasizes quickly based on nothing out of the ordinary other than one more grain of sand dropping on too large a pile. Idiosyncratic, small variables matter. Luck can be more useful than effort. Each situation is path dependent and even a tactic that works during one's first year of programming might begin to fail by the third year, as opponents wise up and change their approach.

But there are several ways in which change can be encouraged and measured in this environment.

Emergent Behavior: Change the Rules, Change the System

The first clue for finding a path for change in complex systems is their tendency toward emergent behavior. In all complex systems, individuals, acting on their own with no central direction, tend to follow a series of simple rules of thumb that direct a good percentage of their behavior within the system. These rules of thumb that emerge from each actor responding to the incentives of the system provide a key for creating and measuring change. Change the incentives or system structure, and one can change a great deal of individual behavior that emerges from those rules.

The most common example of emergent behavior is birds flying in a V formation. The birds don't know they are forming a V, they don't need to know where all the other birds are, and they don't need a map to be constantly, dynamically updated with each turn—as central planning or vast aid strategies to create change from outside the system would require. Instead, each bird follows a few simple rules about where to position itself vis-à-vis the next bird. All it needs to do is keep those rules in mind, and no matter where the flock as a whole is going, the bird will be doing its part to maintain the aerodynamic V.

Emergent behaviors where many individuals self-organize their actions based not on an order from on high, but by following the incentives of a system or the social rules of their group, are everywhere in human social life.

Professors, Politicians, and Emergent Behavior

Emergent behavior based on a few simple rules is generally informal and, with a little pattern spotting, it is not too hard to discern.

In academia, for instance, such rules are based on the key bottleneck of the system: tenure. To make tenure at most universities, one must: (1) amass a large number of publications in peer-reviewed journals, (2) teach in ways that garner good student reviews, (3) maintain good relations with the group of academic peers who must decide whether to accept you into their guild. Thus, junior academics focus on publishing vast amounts in journals that may barely be read by the general public. They often focus on smaller questions to enable more frequent publications: answering a big, hard question takes more time. They learn early on not to grade too hard at competitive schools where students expect good grades, fearing bad reviews. They engage in questions that interest their academic peers, rather than those being debated in the public sphere. By the time an academic gets tenure, these behaviors have been habituated for years, and may be difficult to

alter. Trying to change each of these seemingly separate behaviors is like tackling a hydra. But altering the criteria for tenure could affect all of them in one fell swoop.

In politics the world over, the rules usually stem from two bottlenecks of that system: campaign financing and votes. Politicians respond to the needs of those who can finance or vote for them—and they tend to be rather insensitive to other constituencies. A highly conservative politician suddenly placed in a situation where he must appeal to voters on the left and the right tends to moderate his votes. New rules that limited party funding from any single donor in the United States reduced the value of ultrawealthy businesses and raised the influence of lawyers and other upper-middle-class bundlers who could bring together many maximum contributors. This is also why there is frequent contention over whether U.S. voters should be permitted to register on the same day that they vote, or to use mail-in ballots: such structural rules allocate power to different demographics likely to vote for different parties. And if you change the rules, you change the whole system.

Rules are often incentive based, but they also respond to social behavior, because people exist not as isolated individuals, but in social systems. Corruption is very rare in Chile, despite low public service salaries, and quite high in neighboring Argentina. Once corruption becomes normalized, as it has in Argentina, the social pressure to engage in corruption, or turn a blind eye to one's colleagues' theft, increases. Likewise, in Chile, opportunities for corruption are reduced and the behavior is less common.

Forcing top-down change on complex systems is hard and leads to many side effects: there are too many actors, too many interdependencies, and too many interactions to consider. But affect the rules of the system, and one can affect the behavior that emerges from many individual choices. Measure changes to the rules, and one knows much more about the system than could be gleaned from measuring individual behavior.

Fractals: Patterns Replicate at Large and Small Levels

Because relatively simple rules shape these systems, they also exhibit strong patterns. In fact, patterns in complex systems are so strong that they replicate at multiple levels. In physical systems, the same shapes can be seen if one zooms in close, or looks at the system as a whole. Concrete manifestations of this "fractal" pattern can be seen in the startling, self-similar shape of the Mandelbrot Set, which maps the underlying mathematical expressions of feeding an algorithm back in on itself in a feedback loop.[27] In social systems, similar patterns emerge at bigger and smaller levels. This is not to say that they

can be mathematically plotted, as fractals can be. But it means that spotting a pattern of behavior at one level suggests that it is worth looking for across the system.

This all sounds very abstract, but again, the case of Colombia shows what fractals, or self-similarity at different levels, look like.

Colombia has not only suffered from high levels of political violence, it also has high levels of criminal violence, and of domestic violence. This does not arise by coincidence: these levels are connected. First, high levels of political violence that preoccupy the state provide impunity, making those tempted to use violence more likely to do so. Second, political and drug violence has forced large numbers of families off their rural land. Along with other unemployed men in city slums, some of their sons, often unprepared for work in the city and resentful of their marginalization, provide a cheap pool of men willing to commit the economic violence carried out by drug cartels and urban gangs. Meanwhile, these dispossessed families also appear to have high levels of domestic violence. The stresses of poverty and slum life are compounded by family changes: while women can often find work as domestics, their spouses and sons cannot work in agriculture and often fail to find legitimate work. The changing power dynamics in a patriarchal culture can lead to more domestic violence. The cycle can also go in the other direction: children born in families with high levels of domestic violence frequently re-create that violence in other spheres, in part perhaps because violence occurring at young enough ages alters brain chemistry.[28]

Emergent behaviors where many individuals self-organize their actions based not on an order from on high, but by following the incentives of a system or the social rules of their group, are everywhere in human social life.

Patterns are essential to helping those trying to effect change in complex systems understand the system and design programs for reform. And the fact that patterns tend to be similar at different levels can help observers spot them.

Pattern spotting requires measuring for multiple variables, rather than directing measurement at single variables, however robust those variables are. The presence of patterns also suggests that measurements can usefully be taken at different levels—for instance, that measurements of domestic violence may reveal something about where drug violence may head, and vice versa. Criminologists who have begun to look at violence in public health paradigms—as a "social disease" that is contagious among people and groups—are already seeking ways to use the connections between levels of violence within groups of people to address the problem.

DESIGNING PROGRAMS FOR POLITICAL REFORM

The nature of complex systems suggests a series of principles for designing programs to create change. At its core, the goal is to alter the underlying rules of the political system so that the reform can take place—and so that a coalition of locals remains committed to it and able to fight back when counterreforms are proposed or new leaders are in power. Designing programs that can do this is not easy. But a few rules of thumb can help.

KNOW THE SYSTEM

To affect the rules of the game, one first has to understand them. That means conducting an initial assessment to gain a deep understanding of the system as a whole. This takes time: two-week assessments just won't work. Four to six months should be considered normal to start getting a fingertip feel of social and political systems so that a program designer can begin seeing patterns.

Most important in such an assessment is gaining an understanding of the formal and informal rules of the system.[29] For example, how does someone get power in this country? How are different demographics and political groups aligned with regard to reform, and why? What sorts of behaviors are driven by these rules? This does not need to be a formal "political economy analysis," an approach that was intended to do just this, but has

unfortunately grown into rather academic, 90-page papers somewhat divorced from how they will be used. What is needed is an understanding of the country's dynamics that anyone who grew up there will have intuitively, but that someone from outside must learn by unlearning their assumptions and diving in. A baseline assessment would attempt to understand where the system is starting, what patterns it contains, and whether there are any outliers that might disrupt the system.

Understand the Starting Point

Where is the system starting from in terms of the political dynamics around this reform? Given sensitivity to initial conditions and path dependence, how hard is this reform likely to be?

For instance, in Guinea-Bissau, the generals who won the war of independence in 1974 form the power elite. New political entrants must align with them. A reform program that directly challenges their interests is going to be tough. In most post-Soviet countries, militaries are weak while interior ministries are the real power players; regimes tend to use the police, not the military, to enforce their will. Reformers may be allowed to assist in building modern, civilian-controlled, professional militaries—but they may also find that this does not affect the human rights, transparency, or accountability issues that plague a country and are more based in the interior ministry. The idiosyncrasies of power, elite dynamics, popular ability to press the government—these are all part of where the system is starting from.

Look for Patterns

Are there any patterns that indicate the likely political trajectory of reform? Factors to consider include:

- the strength and intensity of support for and opposition to reform;
- who has power in the system, and the relative weighting of supporters and opponents;
- the size and leanings of the undecided;
- the rules or history of how political parties interact;
- the rules or ways citizens interact with their politicians;

- the demographics of each of these groups projected into the next decade and what that means for likely political positioning on the issue;
- where the major media stands on the issue and how intensely it is covered;
- the rhetorical shape of the public debate.

The data can be both qualitative and quantitative. However, based on the acknowledgment that the program is political and policy uptake is likely to be influenced by political factors, it should be much broader than the data normally collected on a development program.

For example, a traditional development indicator may simply list the immunization rate in Country A. Donors can compare the rate to those in peer countries, and, having determined that they are lower, look to implement an immunization program in Country A.

Pattern spotting requires looking at the indicator in a broader context. For example, in one scenario, healthcare provision is crumbling, and both parties are being blamed as one controls the state level and the other the national. The parties have an incentive to cooperate on improving health outcomes, and each wants to take credit. Parties in this country have a history of elite agreement and then top-down corralling of their junior members, and elites are in agreement about adopting evidence-based immunization programs, with each vying to outdo the other on implementation. Public trust in state institutions is moderate, and institutions have a low to moderate level of corruption. This suggests that a relatively technical immunization program is likely to move forward in fairly linear fashion, with multiple parties in agreement.

That is a very different pattern from one in which the healthcare system is falling apart, and the national government that controls nationalized healthcare is being blamed. In this scenario, the opposition has made immunization programs a wedge issue and is calling for populist policies. However, the opposition has alienated nurses and doctors, who are campaigning for the government. While healthcare workers are seen as corrupt and untrustworthy by the majority of the population, they resent this status and see themselves as overworked, underpaid, and unappreciated. This is the pattern of an issue that is likely to swing back and forth, with reforms and counterreforms. Program designers should be aware that immunization project methods are likely to be chosen based on political point scoring, photo ops, and sound bites rather than sound policy. Evaluators should be prepared for reform outcomes that may vary widely depending on the outcome of elections.

A third pattern might be one in which the healthcare system is falling apart, in part because religious leaders have started to target doctors and nurses as Western imports who are sickening children. Doctors and nurses are refusing to enter rural areas where these beliefs are prevalent. Politicians seeking the traditional vote have sided with immunization skeptics, forcing politicians who have backed a major World Bank immunization

program on the defensive. This is a climate in which a best-practice immunization program is unlikely to function at all. Program designers may need to find ways to change attitudes and provide security to healthcare workers before any immunization program can begin to show results.

Be Aware of Potential Disruptors

While outliers, by their nature, can't be predicted, they can be assigned greater or lower levels of probability. They can also be scanned for indicators that the system is teetering on the edge of a tipping point for or against reform. Simply undertaking the process of becoming aware of such potential changes—a coup, a major corruption scandal, an insurgency, a government getting voted out of power—will allow more realism in judging reform success.

Are any outliers or external shocks starting to appear that could disrupt the system? Could these affect reform negatively, and how likely are they? Could they provide a window of opportunity? For instance, are there political entrepreneurs, a new party, or highly visible individuals pressing for change? Is some portion of the press clamoring on an issue? Is there a nascent social movement forming around or against reform? Is the economy teetering? Is a political scandal starting to bubble?

FOCUS ON INTERVENTIONS THAT ALTER THE RULES OF THE GAME

The deepest, most effective reforms are those that alter the rules of the game. In designing programs for change, reformers should look for bottlenecks and other rules that shape the behavior of many, and work to change them.

New Rules of the Game

In Indonesia, the dominant political parties were long in cahoots with various corrupt private groups or the military. Efforts to tackle corruption within the existing party structure were highly constrained, even when an individual reforming politician took the helm. Allowing local elections pushed some corruption down to the local level. However, it also gave local politicians the opportunity to show that they could be effective and not corrupt. One of these local politicians became mayor of a middling town, then governor of Jakarta, and then, in 2014, president of

the country. Those local elections changed the rules of the game. Unfortunately, opposition politicians saw this before the international community did, and immediately ended direct elections for local offices after the new president's victory. The politicians knew how important rules of the game were to prospects for reform, and they wanted to move them backward: this was a counterreform measure that a political analysis would have predicted and for which opposition funding could have been planned. And, in fact, reformers successfully pushed back a few months later and succeeded in enabling direct elections once again.

Similarly, in the gay marriage debate in the United States, a younger demographic has emerged with fundamentally different views on homosexuality than older generations. These differing cultural views were shaped and nurtured by mass media and the greater openness of gays themselves within society. Expanding the political voice of that generation by registering young people to vote, polling their preferences, and taking other similar steps showed politicians the writing on the wall, and began to move the entire U.S. system. This shaping of a generation's views and then providing a megaphone for its voice had more impact more quickly on the entirety of the gay rights debate than any single issue fight. A new, large political demographic alters the rules of the game in a way that few single issue fights can.

Rules of the game often have to do with who decides—who gets to have a voice in the debate. That is why fights over freedom of the media, the need for a broad media landscape that showcases multiple viewpoints, the right to form civil society groups, and the right to organize are so important: they are all paths to build a voice for those outside the current realms of power. It is why organizing is important to overcoming collective action problems and galvanizing new voices with collective strength. Who is invited to meetings also matters: the Open Government Partnership has started to change the international rules of the game by putting civil society and governments on equal footing in an international forum, this one dedicated to transparency and accountability.

It is also important to consider what beliefs or activities are outside the realm of the acceptable and how they are punished. For instance, can people participate in a gay rights parade without physical fear, and will the police protect them or stand by if they are beaten? Can villagers organize on behalf of a cleaner environment without fear of arrest?

Providing information to alter public opinion, strengthening forms of accountability to reduce impunity for violence, and providing greater political rights to speak and organize can all alter the rules of the game for what is acceptable public discourse. This list is not exhaustive, but it begins to give a sense of where to look for pressure points that can alter a system.

ENGAGE LOCAL PARTNERS TO TAKE ON THE MISSION AS THEIR OWN

External funding rarely lasts long enough to weather the inevitable multiple fights that must be battled to attain any goal. Outsiders simply don't have the staying power for long-term political change: only locals do.

Therefore, funders must engage locals who are more passionately dedicated to the cause than the outsiders are, and who have local support. These individuals may have been working toward the goal before outside money became available. Or part of a development program may be to create a cadre of people who care passionately about an issue for nonmonetary reasons; many next generation leaders programs follow this theory of change. In either case, reform needs local agents of change who are likely to continue the struggle through multiple swings of the pendulum.

The deepest, most effective reforms are those that alter the rules of the game.

These locals need broad-based support within their own society, and sometimes, from outside actors—in politics, one must fight power with power, not just with good policy ideas. Strategies based on small NGOs in the capital with technical expertise are often best augmented with organizations that can bring together larger groups with broader, mass movement dynamics, who can continue to press for change. Crucially, when local NGOs are forced to address problems defined by foreign donors (especially when these change every few years) rather than by their own societies, activists can lose touch with popular desires, and therefore lose the ability to amass broad local constituencies. Social change needs a power base that can force the hands of those who are against reform. This means that donors must be careful to allow local agendas to be determinative, and to look for causes and groups that can garner broad-based support—even if they use language that is more populist or not as nuanced as donors usually like to hear.

PREPARE FOR A WINDOW OF OPPORTUNITY BEFORE ONE OPENS

Most funders look for incremental movement toward goals, and spending that does not seem to result in impact can be written off as a waste. But the reality of political change is that critical junctures, or windows of opportunity, are unpredictable.

When the window opens, however, it will not stay open long. Reformers must be ready to act when the situation tips or small forces align to create a big change. The World Bank's *Doing Business 2007* report found that "in the top reforming economies in the past 3 years, nearly 85% of reforms took place in the first 15 months of a new government."[30] My own research also points to the importance of the first two years of a new administration for enacting reforms. Other windows of opportunity that galvanize people to the streets or impact a media cycle might last only a few weeks or months. That is a tight timeline to implement new policy projects: the ideas and coalitions to support them must exist before reformers take power.

A new, large political demographic alters the rules of the game in a way that few single issue fights can.

Being prepared with a policy proposal, or having a reform constituency that has been formed over years, will pay dividends in that moment—but it may do nothing right up until then. Providing resources for the development equivalent of R&D and allowing reform organizations to invest in building ideas, coalitions, and the other infrastructure of change—without expecting a payoff within that funding evaluation cycle—is the only way organizations can be ready when a window suddenly opens.

STATE GOALS CLEARLY, BUT MAINTAIN FLEXIBILITY AND EXPECT PROGRAMS TO BE ALTERED

Starting with a problem as defined by the society itself, and then generating a theory of change for how to address the problem, is still crucial. It is useful to articulate the problem, determine a goal, and create a strategy to affect it, as the dominant logic-frame thinking forces organizations to do. As the Rockefeller Foundation's Zia Khan said in the summer of 2014, "You learn more by being specifically wrong than by being vaguely right, and in many ways, an initial strategy sets a wheel of ongoing learning and adaptation in motion."[31]

The problem with logic-frame analysis lies not in these important analytical steps, but with rigidity in sticking to a given strategy, or even a given metric, as facts on the ground change. Henry Mintzberg, the well-known management thinker, said over a decade ago, "you don't plan a strategy; you learn a strategy."[32]

Both reformers and funders should therefore begin with a problem they are trying to solve, a theory of change, and impact goals—but not with a rigid multiyear set of

activities, objectives, and metrics for analysis. Instead, they should simply define a first set of steps. After that, programs must be subject to testing, and programmers should expect and encourage alterations as they learn more.

Because complex systems are literally unpredictable, and cannot be modeled even with the strongest of computers, the theory of change, the strategy, the program design, and even the metrics initially selected must be treated as living documents: hypotheses to be tested and reworked, rather than goals to be measured against, as the Problem-Driven Iterative Adaptation model developed by Andrews and others suggests.

In current program design, most of the time is spent at the beginning, with large resource tranches provided at regular intervals, and evaluation coming at the end. But in cases where uncertainty is high and the one certainty is that design is likely to evolve, time and resources are not best spent this way. While more initial assessment can help guide choices, no amount of up-front analysis is enough to be certain that one's assumptions are right. In the words of Horst W. J. Rittel and Melvin M. Webber, "One cannot first understand, then solve."[33] Instead, one must gain some understanding, act, test, gain additional understanding, act, and test again, in an ongoing process throughout the life of a reform.

For that reason, a "spiral development" format, like that recommended by PDIA or used for the iterated design of some material goods, is the best approach. Early iterations are quickly produced and tested under relatively low-cost circumstances to provide greater information, and ideas are rapidly discarded when they fail.[34] Program designers should not assume they know the right strategy from day one, or that past tactics, used in a slightly different historical or social moment, will work again. Instead, it is better to create experimental designs that place a number of bets on different theories of change or strategies, each of which generates further information on the rules of the system to allow for refining programs and improving pattern spotting.[35]

Flexibility is needed not just in program design, but in budgeting, so that resources can be put where they are most needed at a given moment, not where they were expected to be useful three years earlier when program documents were first crafted. Sensitivity to the political economy of donor systems is also required. For instance, if different parts of a program are given to different contractors, they all have incentives to declare that their programs are working. To be able to easily close down portions of a program that are not working, a single contractor will likely need to manage multiple program lines, with the prior agreement and expectation that some areas will be stopped as the program continues, and that such choices are part of success—not an indication of failure.

USE PROGRAMMING TO TEST HYPOTHESES

Allowing flexibility and alteration in programs is not enough. It is better to alter programs based on hypothesis testing, which adds rigor to the complexity of political problems. If program designers start with a description of the problem and a theory of change, programming can be accompanied by clear statements of one's assumptions, so that one can be equally clear if they turn out to be false. Designing the initial steps of a program then becomes a set of if-then propositions, based on whether the hypothesis is proven correct or incorrect.

For instance, an assessment of a country may find that the president is consolidating power in the executive. To understand whether this is a positive opportunity or a negative trend, a program designer could compose two hypotheses, and then develop programs that would test them.

Flexibility is needed not just in program design, but in budgeting, so that resources can be put where they are most needed at a given moment, not where they were expected to be useful three years earlier when program documents were first crafted.

Under one hypothesis, the country's new president is committed to keeping his campaign promise to reinstitute order and security, using every means at his disposal. The government is therefore consolidating power and control over institutions such as the supreme court, the attorney general's office, and the intelligence agencies, and using the military as its main tool for law enforcement in order to bypass corrupt and ineffective police and court institutions. The president is building his own power base and gaining total control over his party and other political institutions so that he can fight other powerful elite factions and improve security. While the choice to rely on the military rather than civilian law enforcement may be problematic, the military is clearly better trained, equipped, disciplined, and prepared than the police. Moreover, it maintains the highest level of public trust of any state institution in the country.

Alternatively, the president may be consolidating power and control over security and justice institutions in order to pressure opponents in other elite groups; control the distribution of economic rents and state-based economic opportunities; and selectively enforce laws for personal benefit and the benefit of those in his circle. By ensuring that the courts and military are beholden to him and undertaking massive police purges, the president can reward supporters with government contracts and patronage positions, punish enemies through selective law enforcement, and use wiretapping authority to ensure

control over the other elites. The president may choose to make deals and agreements with other elite factions to bolster his position and increase his base of support, continuing the basic structure of political instrumentalization of the state.

To test these hypotheses, a program could look for initiatives that would provide data on which scenario was more likely. For example, reformers could support an independent body to track homicide statistics. If the president supported the organization, it would provide evidence for the first hypothesis. If the president wanted a subordinate statistical body, it would provide evidence for the second. With very little funding expended, program designers would have one important data point about the direction of the country. They could then decide whether it made sense to invest more with government organizations, or in accountability institutions that could provide checks on power.

Crucially, there has to be no punishment for getting things wrong—testing is the whole point of putting forth a hypothesis. If program designers are subtly rebuked, lose prestige, or their programs are judged poorly when hypotheses are incorrect, then the hypothesis-driven model will be useless.

PLAN FOR SECOND ACTS AND ONGOING COALITIONS

One certainty of political change is that reformers are working against opponents. The losers from any reform are likely to regroup and try to pick another battle to regain lost ground. Or, if their primary path is blocked, they may redirect action to get at their goals through a secondary path.

As an initial strategy is tried, opponents will align against it, and the landscape of reform will change. The next stage in the battle will take place with the opponents of reform knowing the initial moves and tactics of those pushing for change. A new push for reform may be necessary even at the end of a successful reform, as that is when unintended consequences and side effects are likely to start becoming apparent. Mitigating the worst potential countermoves or unintended consequences of reform means that even as a reform effort is celebrating and winding down, programming may need to be altered—or a new tranche of programming may be needed to address the gathering storm.

For example, in Chile, funding for the NGO sector began to dry up as soon as the Pinochet dictatorship was ended, just as would later occur in Georgia after the Rose Revolution. In Romania, the push for justice reforms declined as soon as the country was admitted into the European Union. But this was precisely when counterreformers were gathering their ammunition in Chile, when reformers in government were changing stripe in Georgia, and when the reformist minister of justice was thrown out of the

government in Romania. After a short honeymoon, efforts to roll back reforms start, and too often, those who fought the first round have disbanded, are exhausted, or lack funding to continue the fight.

If political reform were not seen as a single push, but as a series of ongoing battles, then programs would routinely include funding to retain coalitions after the initial win.

If political reform were not seen as a single push, but as a series of ongoing battles, then programs would routinely include funding to retain coalitions after the initial win. Programs would be designed and funded so that reformers were not exhausted after the first fight, but had the energy and support to push back against the inevitable counterreform. And evaluators would look to the sustainability of the reforming group after the initial fight as one indicator of success.

MEASURING PROGRAM SUCCESS

The development community has, for good reason, taken up the idea of measuring results with fervor. When done well, empirical measurement has overturned assumptions and shown what actually works in a host of areas involving individual behavior and the delivery of services, from getting people to vote to helping them take their medicine. The gradual accretion of small bits of evidence from many disciplines has led to better interventions in multiple arenas. The development field has benefited from such measurement, from the work of the Poverty Action Lab at the Massachusetts Institute of Technology (MIT) to empirical studies of crime control. The idea that strategies for delivering government services or changing citizens' behavior can be measured, and that such metrics, if well chosen, can improve efficiency and uptake, is useful and largely accepted.

And therein lies the rub: the growing movement toward metrics is teetering on the edge of deifying two particular methods of measurement that are excellent for assessing what to do when reformers are in power and able to pick their policy options—but are both unsuited for assessing how to get such change adopted in more contested political and social settings.

Rhetorically, the development field has begun creating a hierarchy with randomized controlled trials at the top, followed by other quantitative methods, generally involving

some form of multivariate regression, with qualitative methods on the bottom. In practice, the worst form of fast and cheap qualitative assessments often prevails, using a methodology that starts with a two-week study tour to assess a problem, then specifies a design for a one- to three-year program based on a single theory of change, logic frame, action items, and indicators of success. These methods of measurement are profoundly unsuited to studying social and political reform. (While decrying the common two-week assessment structure, I've been forced to undertake them myself, fighting along the margins for a few extra days in-country, and I speak from experience about what a poor form of evaluation they are.)

This is not an argument against empirical measurement. Empirical doesn't imply any particular methodology, quantitative or qualitative: it simply means data derived from observation, rather than deduced from theory. The field needs rigorous, empirical measurement that provides more evidence-based understanding of how political processes work and how reform actually occurs in different settings. The problem with the development field's current explosion of empirical studies of service delivery is not with empiricism—it is that the particular empirical methods used tend to be RCTs and regressions that don't allow these studies to answer the question: how can reformers get a goal implemented through a political process?

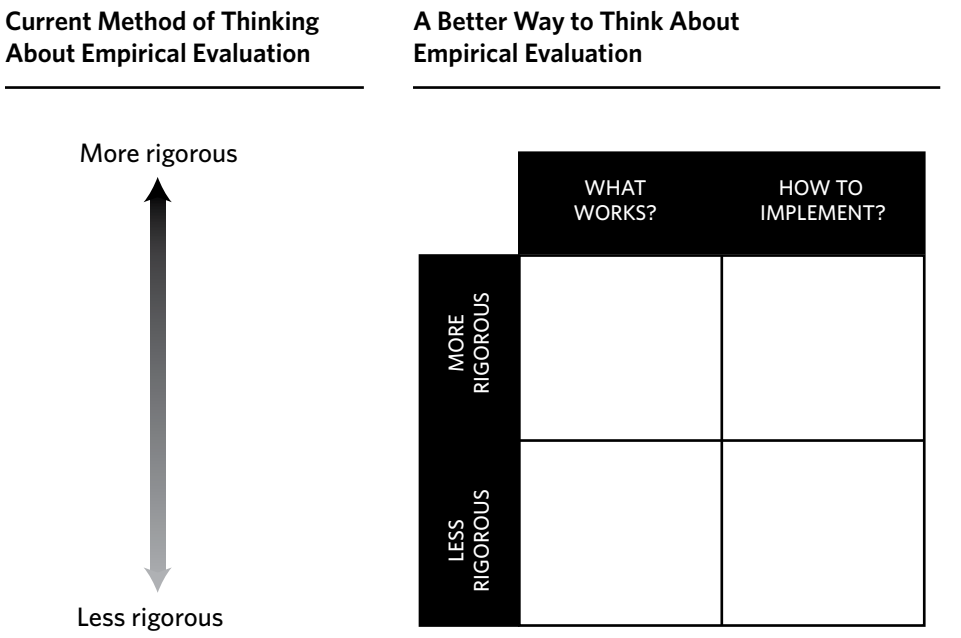
This is also not an argument against quantitative measurement. The fight between quantitative and qualitative data is old and not particularly interesting: both have their value, and both have a tendency to be misused. On the qualitative side, anecdotes and expert status combined with quick seat-of-the-pants impressions often substitute for rigorous data. These forms of assessment are atrocious.[36] On the quantitative side, poor numbers based on badly designed trials too often lead to misleading results that are overgeneralized.

Rigorous but mixed methods are often the most appropriate, when possible. As Chris Roche and Linda Kelly explain in their Developmental Leadership Program publications, the use of mixed methods allows one approach to compensate for weaknesses in others; lets designers triangulate findings, explore different elements, and uncover paradoxes and contradictions; and offers context—altogether providing greater rigor than any one study type.[37] In other words, the development world needs to shift from a universal hierarchy to one that looks for high rigor and empirical evaluations that are suited to the topic being measured (see figure 1).

This last point needs to be underlined. Measurement needs to align with an understanding of how the world being measured actually works.[38] If one believes that the world of political reform involves many actors whose interactions are interdependent, and that creating change is highly path dependent, then measurement must reflect that understanding.

FIGURE 1

MAKING MEASUREMENT WORK FOR REFORM



Instead, development practitioners and donors have somehow ended up, in effect, using a tape measure to determine the amount of water in a pond. There is nothing inherently wrong with multivariate regression, just as a tape measure is a perfectly good tool. But just as the latter is not ideally suited to determining volume, the former is not constructed for a world of interdependent causation, or of multiple causation in which many variables need to interact together to achieve an effect.[39] Nor do many current quantitative studies take time, path dependence, and iterated action into account, as game theory and complexity theory both suggest are crucial for understanding political decisionmaking and reform. As Peter A. Hall writes, "Theories of strategic interaction and path dependence both see the world not as a terrain marked by the operation of timeless causal regularities, but as a branching tree whose tips represent the outcomes of events that unfold over time."[40] Randomized controlled trials are excellent at showing the outcome of a given intervention. They are extremely poor at showing the process that produced the end goal, or determining whether that process can be replicated.

For instance, a very interesting forthcoming review tests the role of wages, incentives, and audits on tax inspector corruption in Pakistan. But even once these findings are

The development world needs to shift from a universal hierarchy to one that looks for high rigor and empirical evaluations that are suited to the topic being measured

determined, the study will not explain how to get the best mix of wages, incentives, and audits enacted into policy—much less implemented once they are declared.[41] In fact, the best researchers know this. Rachel Glennerster, executive director of MIT's Poverty Action Lab, is open about the fact that RCTs have taught the development world a lot about what works, but she acknowledges that it is difficult to draw general lessons about implementation from these trials.[42] Or, as Angus Deaton suggests, RCTs are "useful for obtaining a convincing estimate of the average effect of a program or project, but the price for this success is a focus that is too narrow and too local to tell us 'what works' in development; to design policy, or to advance scientific knowledge about development processes."[43] Too often, RCTs are rigorous because of the care in statistical design and testing, but they exhibit a lack of rigor in extracting policy-relevant messages from that evidence base.

In other words, much good empirical work has been done to answer tough questions about how to create the most efficient and effective program—once that policy has been decided upon and agreed to by political actors. But they are fundamentally different questions from how to get such a reform agenda enacted into policy in the first place in a political world, and how to measure the success of that effort.

PITFALLS OF POOR MEASUREMENT

The use of traditional measurement techniques that look for incremental, linear change based on predetermined logical frameworks—when a program is, in fact, facing a political issue—has led to three unintentional but serious problems: measurements that inadvertently favor autocracies, enclave projects that can be sealed off from politics but not scaled, and the advancement of programs that will look good up to a point—and then face implosion. These pitfalls are forced by the logic of the system, not by a lack of understanding.

Praising Autocracies

As discussed above, in complex reforms involving politics, change can't be measured linearly. The exception is in autocratic systems. This could be why some development practitioners find their measurement systems leading them to rank autocracies as more successful than democracies in reform efforts. Current praise for Rwanda, for instance,

which ranks 32nd in the World Bank's *Doing Business 2014* index, echoes past praise for Ghana under coup leader Jerry Rawlings.[44] The World Bank's *Fighting Corruption in Public Services* report on Georgia rightly praises revolutionary government reforms—but does not mention unintended consequences that the measurement was not designed to capture, such as the reformist government's tightening grip on the media, judiciary, business community, and other avenues of oversight and accountability.[45]

Traditional incremental and linear railroad-style metrics work in countries where the leader sets the end goal definitively, social opinion is highly constrained, and politics can be suppressed (at least for some time, though rarely indefinitely) in favor of technocracy. This means that traditional development measurements sometimes find themselves crediting what some consider to be "benevolent" nondemocratic regimes with making rapid progress. Meanwhile, more open societies appear to be faltering in chaotic processes that stray from best practices and compromise away important parts of good policy. For those who see dictatorship itself as having inherent developmental pitfalls, measurements that put these governments toward the top should also be viewed as having inherent problems.

Providing Charity

A second pitfall to using traditional measurement to gauge political change is that it leads donors to work in technocratic enclaves that don't affect the larger system. By isolating their development projects from politics as best they can, these programs provide beneficial services—but they serve as Band-Aids, addressing immediate needs with external resources rather than catalyzing the country to solve its own problems. For example, donor agencies spend billions on palliative service delivery to shore up healthcare systems, provide electricity, educate students, and otherwise take on jobs that the government is not doing. These projects may build an indigenous skill base and certainly provide valuable services to those in immediate need. They also give the central government the option of allowing donors to focus on helping their people while the leaders focus on stealing assets, helping their ethnic group, or benefitting their home regions instead of building their countries. Programs to assist local champions often act as enclaves that fail when they are confronted by politics: for instance, in Guatemala, donors flocked to a promising attorney general who was effectively tackling violent crime and reducing impunity. Then, in 2014, she was removed by political powers who did not want her to hold human rights abusers accountable.[46]

Development practitioners are used to seeing metrics showing that they have helped to build 100 well-functioning schools: a valuable goal, but a drop in the bucket when 10,000 schools in that district alone are broken. They are happy to make healthcare in one city of 150,000 better—when the country's 10 million other people lack basic vaccines. Often

these programs are called pilots to imply that they are supposed to be picked up and scaled by the government—even when everyone involved knows such pickup and scaling is unlikely to happen. These programs are valuable and those who do them are doing worthwhile work that makes the world better. But they are charity, not development.

Project Implosion

Finally, donor projects that are chosen because they can be easily measured can be actively detrimental. The U.S. government strategy in Afghanistan was to extend the writ of the central government of Afghanistan. However, as U.S. Institute of Peace analyst Frances Z. Brown wrote, "Especially in some rural areas targeted by … aid during the surge, government presence is often viewed as foreign and extortive. Intrusion of government could actually fuel instability to a degree that no amount of goods proffered could outweigh."[47] Similarly, the U.S. military spent billions of dollars to train and equip soldiers in Mali, where they eventually launched a coup against their government; in Iraq, where they dissolved when faced with an insurgency; and in Libya, where their equipment was stolen by militias and is now used against U.S. government aims.

In all these cases, the selection of programs that could be finished quickly and easily and then linearly measured by outputs (buildings built, equipment provided, trainings offered), rather than harder-to-measure factors, yielded efforts that backfired in their ultimate goals. Many of the practitioners involved knew this would happen. But they were stuck within a system that wanted projects that could be measured in easy-to-show, numerical formats, for political leaders demanding progress in six months.

So what are better methods to measure success? How can reformers or funders know if their programs are working in such a politically fraught world?

SHAPING THE SPACE OF THE POSSIBLE

Complexity theory again provides a useful answer. It acknowledges that these systems are mathematically unpredictable. But one can still get a sense of what they may do. Thanks to emergent behavior based on simple rules and the similar behavior that takes place at multiple scales, these systems have strong patterns. Complex systems also exhibit what is known as "dynamic stability," undergoing constant change while maintaining a basic structure.

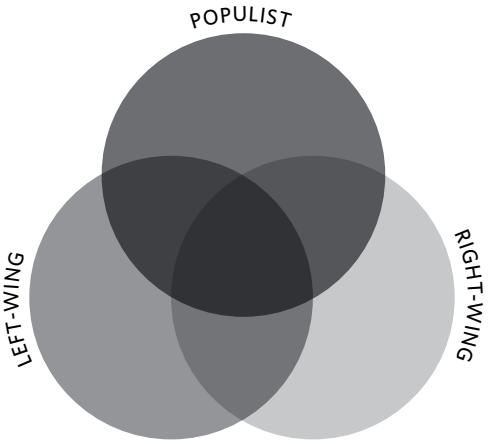
So while it is impossible to predict where these systems will go next, one can map the contours of possible outcomes. And, to measure the success of a reform effort, one must consider whether that space of what is possible has changed.

Consider a dripping water faucet: it turns out that it is impossible to predict exactly where the next drop will fall from along the circumference of the faucet. But it is very possible to draw a circle around the space within the sink where all the drops could hit. That area is known as "phase space"—the space of the possible. Because of the similarity and patterns that underlie complex systems, they all have a phase space within which activity takes place.

Phase space does not always look like a single circle under the faucet. I've talked about how political change often tips suddenly from reformers to counterreformers. This all happens within the space of the possible—jumping precipitously between two sets of policies while maintaining the basic structure of the system. It is as if water is dropping from one half of the faucet's circumference for a while, then suddenly shifts to mainly dropping from the other half. In fact, this sort of jump between different parts of phase space is common in complex systems. These systems often have multiple equilibria each with its own attractor that pulls activity toward itself. So phase spaces may look like a Venn diagram with three main attractive points among which activity jumps (for instance, politicians favoring left-wing, right-wing, and populist policies), or like a figure 8, with a lot of activity gravitating around two points, or take other configurations, depending on the political structure involved (see figure 2).

Picture U.S. policy flipping between approaches favored by Democrats and those preferred by Republicans—policies may differ quite a bit depending on who wins an election. These are the two poles of equilibria in the United States, each of which attracts policy that is more left- or right-wing. The policy phase space encompasses them both,

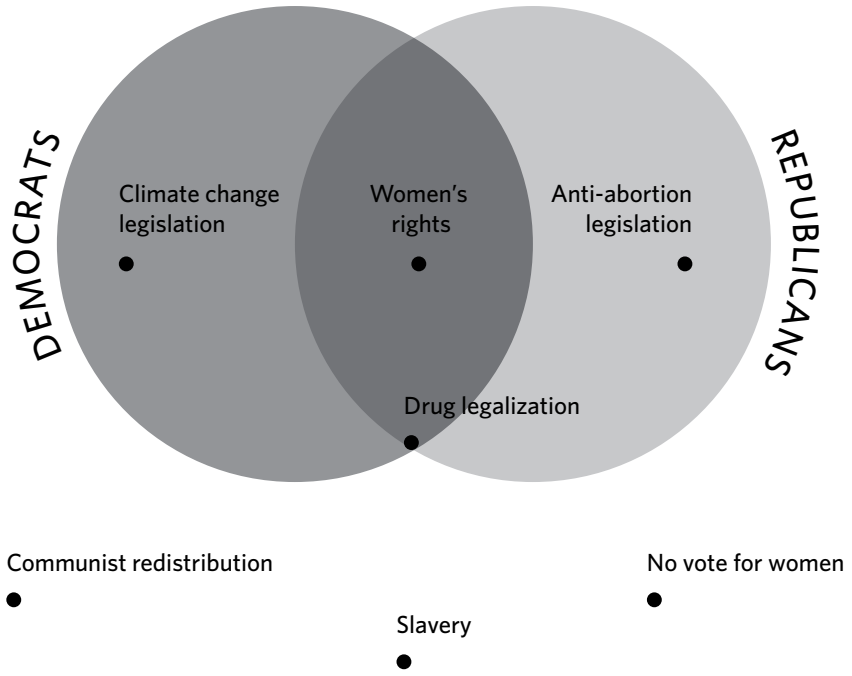
FIGURE 2
AN EXAMPLE OF POLITICAL PHASE SPACE



and jumps between the two (see figure 3). Yet phase space still has a shape and coherence: truly out-of-the-box thoughts just don't get a hearing. The phase space of U.S. politics currently encompasses Republican and Democratic ideas of a certain stripe, but Leninist redistribution, slavery, disenfranchisement of women, and universal government-provided healthcare are outside the U.S. political phase space—these policies are just not going to be broached under the current political configuration. Yet in past eras, some of these issues were within the phase space—and in the future, others may enter the space of the possible.

FIGURE 3

PHASE SPACE OF U.S. POLITICS IN 2014



The jumping among multiple equilibria within the same phase space can make determining success difficult, as the case of gay marriage showed. Measuring the progress of reform at one moment in time can create a wildly different impression than measuring just a short time later. Assuming that you are mapping the whole space of the possible, when in fact you are just looking at one piece of the figure 8 or Venn diagram, is what leads to massively wrong predictions—such as assuming linear policy change just before a revolution.

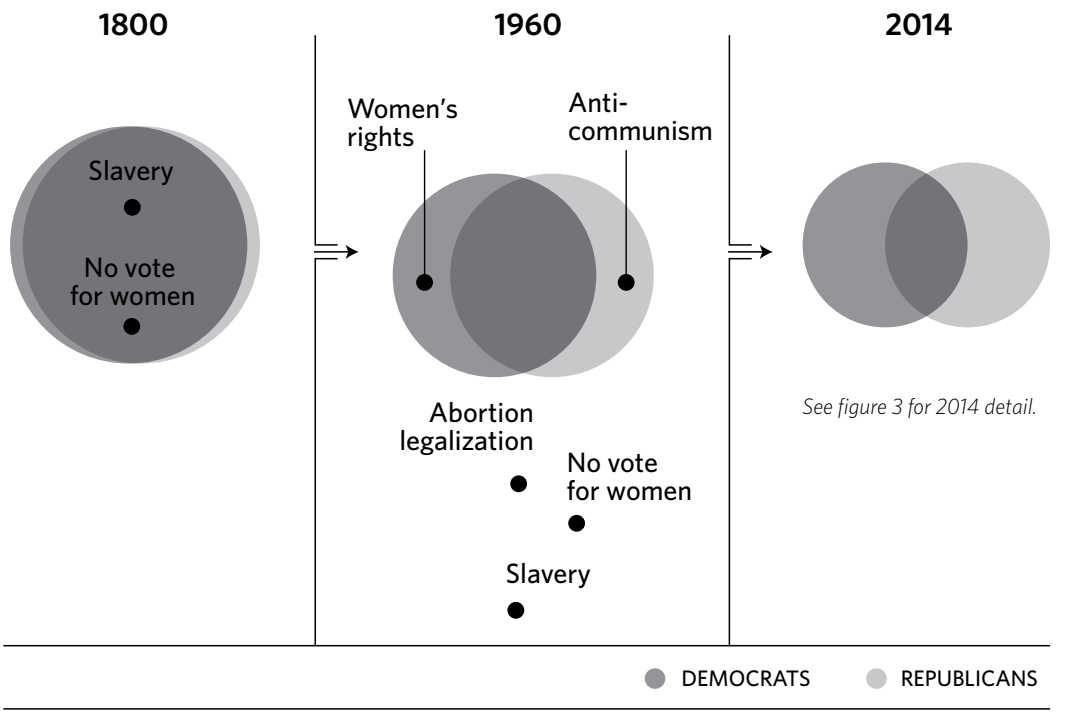
Determining success, therefore, requires broadening the picture and measuring the entirety of the phase space—the realm of the possible options, whether they are present at that moment or not (see figure 4). Measurement must consider not where one happens to be at a given moment in phase space—but what the shape of that space of the possible as a whole is, and whether it is changing.

For instance, install a new faucet with a smaller circumference, and while drips may still fall on the right or the left, the phase space as a whole will be smaller. Push the faucet to the right, and the whole phase space will move. To measure success, one must measure whether the space of the politically possible is moving in a direction that is more amenable to reform and inimical to those who oppose one's policy goals. What was thinkable and discussable at one period in time is no longer on the table in another. That is when one knows phase space has changed.

In the early 1990s, Colombia was the world's most violent country. But the end of Pablo Escobar's war against the state in 1991, followed by the destruction of the Medellin drug cartel and Escobar's death in 1993, led to a sharp drop in homicides. However, as

FIGURE 4

THE EVOLVING PHASE SPACE OF U.S. POLITICS



drug-related paramilitaries grew and fought guerillas, murders crept back up. They didn't get quite as high as they had been: Escobar was a particularly violent character. But they rose because many of the rules of the game still held—drug money was still worth fighting over; the state still sided with violent actors who helped officials' political campaigns and acted against leftists; and there was still a great deal of impunity—the players just had to regroup.

To measure success, one must measure whether the space of the politically possible is moving in a direction that is more amenable to reform and inimical to those who oppose one's policy goals.

Then, from 2002 to 2008, Medellin witnessed a fall in its murder rate so dramatic that it was known as the Medellin miracle. It was attributed to many causes, primarily alterations to urban infrastructure. But in 2008, the homicide rate jumped again, though it didn't get quite as high as it had been in 2002. The shift had a lot to do with one man: when the United States insisted that Colombia extradite the leader of one of the strongest neoparamilitary groups, fights over turf broke out again between the remaining criminal groups, sending the homicide rate back up.

But, crucially, despite the back-and-forth nature of homicide in Colombia, the phase space itself has been contracting since the early 1990s. Some very important changes to the rules of the game had begun with constitutional reforms in 1991. Although they took a few years to make themselves felt, these changes enabled new middle-class political coalitions that, by the early 2000s, had beaten the oligarchs' duopoly control in Medellin and Bogotá. U.S. pressure and military and police aid had reduced impunity and forced the government to clean up its politics to some extent. As the state has become somewhat more effective, and the political system has become more open to a greater variety of political voices, guerrillas have lost the soft support they enjoyed among social democratic–style leftists. Both peaks of violence since 1993 have been lower than the earlier peaks; fighting groups are smaller, and political support for them has declined. For example, after the 2008 leap, the violence in Medellin began to decline again in 2009.

Violence is far from gone in Colombia: the successor paramilitary groups are particularly worrisome. But while Colombia is likely to continue jumping between two different equilibria of higher and lower homicide levels for some time, as long as the rules of the game continue to follow the trajectory they've been on for the past twenty-five years, the overall trend in homicides is likely to continue to decline as impunity and the space of the possible for violent activity narrows. It is easiest to see phase space changes like these when one looks backward. When in the midst of reform, tracking whether the space of the possible is on the side of reform requires looking for these historical trends, as well as

spotting patterns in the opinions of politically important demographics, trends in the size and composition of those demographics, and changes to the rules of the game that may empower new groups and alter who gets to decide. All of these form the basis for better assessment of political reforms.

What does this mean practically? How can development practitioners avoid the pitfalls of traditional linear measurement and instead measure the phase space of reforms—within the realities that most development organizations must work?

An excellent measurement would take the temperature of a program at three useful points: the moment of program design (before taking on a program), midstream, and after funding is complete. In an ideal world, long-term assessments would also take place—but perhaps the best one can hope for in the real world is that the final assessment takes place a few years after program completion, to capture the likely back-and-forth effects.

PROGRAM DESIGN ASSESSMENT: IS IT EXPERIMENTAL AND ITERATIVE?

In the design phase, most development programs are assessed for robustness against a series of criteria, from environmental impact to inclusiveness. For programs that are in some part political, assessment can be based on whether they are structured in ways that are most likely to affect complex systems. A good design assessment would include:

- Does the initial assessment allow enough time on the ground to gain a deep understanding of the system as a whole?
- Does the program design grow from the assessment of the country—and do the program design, timeline, and expected goals seem realistic given the baseline assessment?
- Is the program designed to affect structural aspects of the system?
- Does the program engage local actors who were already imbued with the mission before donor funding began?
- Does the program structure its initial programming to test explicit hypotheses and assumptions and alter programs based on those tests?
- Does the project allow for a flexible, iterated process, so that programs can be altered in response to critical junctures and windows of opportunity?
- Does the program design allow for a multiyear timeline, including follow-up after the date of the expected reform, if one exists?

Initial assessment of a program should, ideally, guide whether a reform effort takes place at all, based on the findings from the country assessment. But because such decisions are often made above the level of program designers, assessment should begin by allowing for the fact that some environments are much harder to change than others, through no fault of the reformers or the reform project.

This is an acknowledgment, basically, of path dependence, and it explains why reforms should be assessed by measuring the delta, or amount of change, rather than simply the end state, to ensure goals that are attainable given the systemic realities. When one considers program-end assessment measures, this delta, not just impact, is what matters.

Based on these factors, those measuring a program could engage in an assessment that borrows from Bayesian statistics. Under this approach, instead of simply stating "this is how the world is," one states prior beliefs about what the world is like, and assigns them probabilities of certainty. Given that level of certainty about the assessment of a country, a hypothesis regarding reform outcomes is created and a probability is assigned.

For some projects, assessors might decide that any reform at this time is not likely to bear fruit. In that case, the project could be stopped or redirected before it gets off the ground. Or the outcomes expected from the program could be made much more modest in a more difficult environment, so that midstream and end-state assessments can be realistic.

MIDSTREAM ASSESSMENT: EARLY OUTCOMES

Because programming is assumed to involve testing among multiple options and design iterations, it makes no sense to measure progress until the program is well under way and has made it through at least one and preferably two rounds of iteration, testing, and alteration based on that feedback.

At this point, metrics should be targeted at program outcomes, rather than impact. These will be highly variable depending on the program. For instance, in a program that uses police training to reduce crime, the impact metric is whether crime declined. But that could be affected by so many factors that it doesn't make sense to measure at such an early juncture, though it should certainly be tracked over time to see whether there is correlation with the training program. Rather than falling back on useless output metrics, such as the number of police trained or the number of trainings held, evaluators should look for useful outcome measurements—such as whether the newly trained police are more skilled in problem-oriented policing after training than before, and whether they are, in fact, using it in their daily work.

Outcome metrics will need to be chosen based on the particular project, and must take into account nonobvious movements in quantitative indicators. For instance, when a police force starts being more effective and gains more legitimacy, crime statistics often rise: people suddenly decide that it is worth reporting crimes that previously went unreported.

Outcome metrics must also take into account the creation of perverse incentives and efforts to game the system—which are inevitable and to be expected in the process of a political or social reform because one is working against opposition. For example, when police learn that they are to be evaluated based on the number of cases closed, some forces have responded by refusing to accept victims' reports for crimes that are harder to solve, for fear that the cases will undermine their statistics.

Because the interdependent nature of complex systems generates constant unintended consequences, and because of the adversarial nature of political change, qualitative research is an important and necessary supplement to quantitative surveys and other assessments. Qualitative methods should be designed to capture unexpected side effects, as well as countermoves and counterreforms that will affect quantitative variables.

Finally, programs can be assessed based on whether they are working through one of the theories of change most likely to affect complex systems, such as:

- Is the process of reform working with a viable and broad social movement or political coalition structured to effect change? Politics takes place through people pushing for their views. Thus, projects that assist local development leaders and politically effective coalitions and catalyze social organization must be a part of the portfolio and are generally more likely to affect the rules of the game than those that provide equipment and infrastructure.[48] As Acemoglu and Robinson found, the most successful reforms are pushed by broad-based coalitions of change, so that one set of rent seekers is not simply replaced with another set.

Building Broad-Based Coalitions

A broad-based reform coalition may look like the progressive movement in the United States in the late 1800s: urban middle-class business, agrarian populists in the Midwest, young idealists, and the Protestant middle class teamed up to support anticorruption reforms, all for their own reasons. Similar broad-based reform movements appear to be forming around environmental issues in China, and in many cases were behind the various Enough! movements in Eastern Europe. In Lagos, business people, market women,

and Western-educated reformers found themselves on the same side as a winning political party pushing for functional taxation. In Chile, both main political parties, the biggest newspaper in the country, elites closely tied to the Pinochet regime, and left-wing human rights activists galvanized around justice reform for a brief period in the early 1990s. Broad-based coalitions do not look like a single NGO in the capital: they engage large constituencies who vote or hold intrinsic power on account of their size, and they are organized around issues that people are talking about on television, radio, or whatever the main form of nonmonitored communication is in that country.

- Is the process of reform shaping the rules of the system to make positive behavior easier, and antireform behavior more difficult? Political reforms should use the reality of emergent behavior to shape the rules of the game so that society itself pushes its members toward the incentivized actions. Several sorts of programs might do this, such as building accountability and oversight capacity; tapping into status, pride, shame, esprit de corps, and peer pressure to alter behavior in ways that are intrinsic rather than pushed from outside; and affecting hiring, promotion, and retention systems.

  In Georgia, the revolutionary government that took power after the Rose Revolution quickly swept away public petty corruption, a reform that has outlasted that administration. To do so, it relied on changing the rules of the game and reinforcing them with status and social pressure.

  Georgia was a country where stealing from the state was not seen as bad behavior when the nation was under Soviet control. Petty corruption had been a way of life for decades, financing many middle-class lifestyles. The new government paid civil servants a living wage, built better office buildings for them to work in, and offered them nicer cars. In a country of widespread unemployment and constant electricity shortages, a living-wage job in a heated, well-lit building was rare and people wanted to keep their positions. That alone, however, would not have been enough: similar infrastructure changes in neighboring countries had no effect on corruption. But the Georgian government also changed the rules of the game, using both hiring and firing power and status. It created and enforced a meritocracy, firing thousands of civil servants who couldn't pass basic skills tests, then explicitly recruited new people to help "change their country." The government also linked their better working conditions with the fact that civil servants were being respected and, in turn, were expected to respect their work.

The meritocracy and status reforms required matching carrots with sticks: those who were caught in corruption were fired quickly and publicly—in front of all their peers, or on television programs where they were hauled out of their homes in the middle of the night, in their nightclothes, in front of the entire country. Suddenly, in a country where connections had been paramount, those who were caught in corruption found themselves publicly stripped of their status, while the meritocracy quickly promoted those who did their jobs well and played by the rules.

Within a year, these reforms had created a sea change in Georgia. Ten years later, with a new government in power, people still expect their public services to be free from petty corruption. Asking for a bribe would be considered shameful to one's social peers, and offering a bribe is just no longer done. Thus, even though the fear of punishment has lessened in a less repressive government, petty corruption remains at bay.

- Is the process of reform creating a small outlier that has the potential to snowball? Outliers matter in complex systems, so creating outliers that can have outsized effects is a fine strategy. If this is one of the project goals, what are the benchmarked measures to see whether the outlier is, in fact, creating a positive feedback loop and snowballing rather than growing incrementally or stalling? A judicial academy might undertake a meritocratic selection process to pick a handful of the highest caliber law students each year, who, over a generation, are expected to infuse the judiciary with pride, status, and an independent outlook. As these judges join the bench, measurements can be taken to see if their attitudes and behaviors are affecting that of their colleagues, or vice versa, and whether the effects of the initial training and meritocratic system dissipate or build.

The point of midstream assessment should be to determine which parts of a program are working, and how to shape future program activities. Looking at programs this way requires a huge change in funding, program design, evaluation, and thinking. Right now, a project that is altered midstream, or has failed to meet preset benchmarks, is seen as an expensive mistake in which sunk costs have been wasted and everyone looks bad. Program implementers and their organizations have strong incentives to avoid ever stopping a program early or admitting that benchmarks have not been met or are no longer useful.

Shifting to an experimental mind-set means that any single project goal (for example, strengthening the healthcare system) may run three or four experimental subprojects within a single program under a single implementer or contractor, and some of these must be expected to end by this midpoint measurement as the project is refined and moves forward. The goal, not the project, is what succeeds or fails.

POST-PROGRAM ASSESSMENT

The funding is over, but the process of social and political change will continue, possibly for decades. What metrics indicate whether this was money well spent? One wants to measure impact at this point. However, in political and social change, impact is a tricky variable. A program may have contributed the final straw that tipped a reform over the edge, and thus may look wildly successful. But the effort could have been poorly planned and implemented: impact is as much about the beginning terrain, luck, and windows of opportunity happening to open up as it is attributable to the program design and implementation. Another effort, starting in tougher terrain, or facing an unpredictable and unfortunate set of external shocks, might have been far better planned and executed, but have no impact to show. If the goal of the assessment is simply to prove impact, this doesn't matter. But if the assessment will be used to determine whether an organization deserves future funding, whether a program designer deserves to take on bigger programs, or whether a strategy is working, these differences matter quite a bit.

Some initiatives that are focused on changing the rules of the game or creating a snowball effect from an outlier might be planning for long-term impact, but may barely have gotten under way when donors conduct their post-program assessment. For example, a program to improve the rule of law may have spent five years creating a training academy for new judges, a meritocratic entry program for police, and an entrepreneurial bar association to press for independence and maintain excellence. But a corrupt government has not yet made many changes to the rule of law. Should the program be deemed unsatisfactory? Its change efforts are generational, and should be measured over decades; judging them before they start to bear fruit just doesn't make sense.

Impact is as much about the beginning terrain, luck, and windows of opportunity happening to open up as it is attributable to the program design and implementation.

Of course, impact will have to be one of the variables measured; it should just not be considered definitive, or even more important than some of the other measurements indicated here, because of its variability and the many factors that are beyond the control of program creators. Moreover, even major impacts can be reversed over time, so patterns of changed phase space, rather than impact itself, must be measured for a more complete view. Additional factors to measure at a program's end might include:

- If the reform has created a coalition, movement, or advocacy group for a specific change, is the coalition growing more robust, powerful, and interconnected? Or is it

consolidating into a few people, and/or is the reform process exhausting the coalition and causing drop-off? Is it creating self-generating momentum and buy-in, such as some local funding, however minimal? The political process requires more than just winning a single battle—judging the directionality and self-sustainability of a movement is an important indicator.

- If formal organizations started by the reform effort shut down, but other spin-off organizations that share the same goals continue to function without external funding, that is a success—not a failure. In Macedonia, for example, an eight-year U.S. Agency for International Development (USAID) project sponsored a business consulting and support service that closed when the funding ended. However, it spun off six institutions offering similar services, at least two of which were still functioning eight years later. That suggests the creation of embedded local interests carrying on the same fight, a path toward significant, ongoing success.[49]

- Has the effort helped coalesce or catalyze a group within the country that is either broad based or highly influential and elite, whether organized into a formal NGO or simply part of an informal community? And have members of that group internalized the value of the reform effort and made it part of their personal missions?

  When speaking to Eastern Europeans who played essential roles in the movement of their countries toward the rule of law and democracy, many cite the Open Society Foundations' high school debating program as crucial to their development. It created a meritocratic, loose community of individuals who wanted to change their region for the better, kept in touch, and helped one another for years. Similar informal groups of change agents in Indonesia, Chile, and elsewhere can be tied to significant reform in their countries over years.

- Has the effort institutionalized any changes to the rules of the game that are being implemented? These could range from expanding the field of who is at the table or who makes decisions (through enhanced freedom of association and speech and stronger independent media) to changing entry and expectations for key professions (such as meritocratic civil service processes or transparent procurement processes to reduce economic capture). It is important to assess whether changes are passed into law or regulation, and equally, if they are being implemented—the former without the latter can be a mark of lip service to foreign donors. If laws are being implemented, that should be captured in quantitative impact assessments of corruption perception, ease of doing business, etc. These should be matched to qualitative assessments of potential side effects, surveying users, not just practitioners themselves.

  In Vietnam, a highly successful USAID program was at first working with the government on building capacity for trade. Those working on the program soon

realized that they needed a whole new way of creating laws that solicited business and other public input. They changed tactics from focusing just on trade to working with the government to build internal support for a "law on laws." When it finally passed, Vietnamese officials credited the change with altering the entire legislative culture of the country.[50] While it may or may not have built trade capacity, it changed the rules of the game for how business and government interacted in a way that will let Vietnamese businesses press for their own needs in the future—a far more important change than simply creating a single new port or road.

- Has the effort institutionalized any social changes that are likely to continue and metastasize? For instance, is there a political party, pressure group, or organized demographic that did not previously exist? Is there a discernible change in public rhetoric? Are attitudes toward an issue changing in a way that is captured in opinion polls or surveys, possibly before such changes are apparent in the political sphere? Are demographic differences trending in a direction that makes social change more possible? Laws and government programs often follow, rather than lead, public opinion; are there long-term alterations in opinion trends that can be measured by looking at baseline surveys, post-project surveys, and the demographics of who holds what views?

  Consider the wildly successful effort to alter food practices in the developed world. Wherever one comes down on the issue, it is hard to dispute that in less than two decades, commercial monoculture farming with widespread pesticide use has gone from being the way nearly all farming was done, and a nonissue to most of the public, to being the subject of a major mainstream public policy debate.

  Europeans have not only banned genetically modified foods in their own countries, they have pushed developing countries whose people are starving to do the same, even when such foods are offered at fire-sale prices. Local ballot measures ask voters whether products should require pesticide and genetically modified organism (GMO) labeling. Demographics are further trending in this direction: beginning in 2001, a series of elite colleges in the United States, from Yale to Stanford, started college farms due to demand from students who were concerned about where their food comes from and wanted to study new forms of agriculture.

  Nor is the issue just class based: mass-market burrito chains such as Chipotle differentiate themselves based on their hormone-free, organic ingredients; organic foods have moved from specialist health food stores to multiple aisles of mainstream supermarkets. This is what a mass movement looks like. Foundations and funders who played a role in catalyzing this change have been immensely successful in altering the system—even if their particular policy goals have not yet come to fruition.

- Has the space of the possible changed, and has the program contributed to this change? This is the big impact question, which can be measured even if the originally desired policy reform has not yet occurred. Answering it could require a mix of quantitative and qualitative methods, such as analyzing public discourse in traditional and social media; considering what policies are being put forward at the end vs. the beginning of the program period; and considering the vertical and horizontal methods of accountability that may now exist to affect government decisionmaking.

Such a series of assessment questions may seem insurmountable, but according to Roche and Kelly, AusAID, Australia's aid program, has implemented a number of these.[51] Its evaluation system includes a basket of indicators such as contextual analysis, flexibility and responsiveness, investment in relationships, commitment to long-term approaches, and support of local leadership processes, with a mix of qualitative, quantitative, and short-, medium-, and long-term measures. This is precisely what is needed to capture political dynamics.

CONCLUSION

Successful efforts in political and social change do not follow the path of a train headed down a track. They are not about traveling a predetermined route toward a preset best-practice goal along a timetable of benchmarks and chalking up incremental victories. Nor can they be measured based only on whether one's policy goal has been met: impact is a slippery variable whose attainment may be as much about luck as skill in the reform effort.

Instead, reform efforts must be adaptive and iterative to test assumptions and counter equally adaptive opponents. Like sailboats, they must use the wind of opportunity when it arises, and expect that they will move sideways at times to get to their end goals. Programs must focus on understanding the underlying structure of the system and the rules of the game so that these rules can be altered to create emergent behavior that is transformative. And programs must be measured based on whether they have affected the space of the possible, and laid the groundwork for long-term, ongoing war, not success in a single battle.

Programs must be measured based on whether they have affected the space of the possible, and laid the groundwork for long-term, ongoing war, not success in a single battle.

Creating new methods to measure social and political change is difficult. But the potential payoff is big: programs that could impact the largest problems of our time, rather than play along the edges, offering temporary charity but not transformation. It has happened before, and it can happen again.

NOTES

1 The connection between more open political systems and more equitable economic institutions that lead to wealth and development has been deeply explored by Daron Acemoglu and James Robinson, *Why Nations Fail* (New York: Crown Business, 2012), and Douglass North, John Wallis, and Barry Weingast throughout their writings, but particularly in *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History* (Cambridge: Cambridge University Press, 2012). Both sets of authors make the case that economic reform toward greater openness and development involves politics.

2 See, for instance, Erin McCandless' work discussing the fights between weak countries and the OECD, World Bank, and UNDP to develop a fragility assessment and a set of indicators of statebuilding and peacebuilding: Erin McCandless, "Wicked Problems in Peacebuilding and Statebuilding: Making Progress in Measuring Progress Through the New Deal," *Global Governance* 19 (2013): 227–48, 234.

3 World Bank, *World Development Report 2011: Conflict, Security and Development* (Washington, DC: World Bank, 2011), 108–109, and box 3.6. Forty-one years is how long it took the fastest reforming countries of the twentieth century to move from average levels of governance typical of a failed state to "good enough."

4 North, Wallis, and Weingast, *Violence and Social Orders*, 27.

5 Lant Pritchett, Michael Woolcock, and Matt Andrews, "Capability Traps? The Mechanisms of Persistent Implementation Failure," Center for Global Development Working Paper 234, 2010, and Lant Pritchett and Frauke de Weijer, "Fragile States Stuck in a Capability Trap?" World Development Report 2011 Background Paper, September 3, 2010.

6 Thomas Carothers and Diane de Gramont, *Development Aid Confronts Politics: The Almost Revolution* (Washington DC: Carnegie Endowment for International Peace, 2013).

7 See, for instance, Anne-Marie Leroy, "Legal Note on Bank Involvement in the Criminal Justice Sector," World Bank, February 9, 2012. Former General Counsel Ibrahim Shihata's guidance regarding World Bank work in areas that had both economic and political effects was that the Bank is required to take into account only economic considerations and avoid entanglement in any local partisan or ideological controversies, seeking consensus rather than choosing sides in issues under debate. In countries where governance may be undertaken by organized criminal groups, narcotraffickers, or elites opposed to development per se, such a stance leads to inevitable program distortions in how issues with a political dimension can be talked about, and therefore how they can be thought through, designed, and measured. Without alterations to the articles, this is inevitable, and Bank staff are doing their best to act within the letter and spirit of the law and their mission. It is also unfortunate.

8 See Matt Andrews, *The Limits of Institutional Reform in Development* (Cambridge: Cambridge University Press, 2013); Matt Andrews and Lana Pritchett, "Escaping Capability Traps Through Problem-Driven Iterative Adaptation," Center for Global Development Working Paper 299, June 22, 2012.

9 Unfortunately, this "wonkwar" was like two ships steaming past one another. I doubt either author intended to suggest that political thinking could only be grounded in "expert" opinion based on anecdotes that are justified by status and seniority, as Chris Whitty and Stefan Dercon accused, or to claim that political thinking did not require empirical evidence, as Rosalind Eyben and Chris Roche unfortunately implied. Political approaches must be based in empirical evidence—that is, observable fact—it must simply be the right type of evidence, and not the technocratic and large-scale forms of measurement that are currently in vogue. See "The Political Implications of Evidence-Based Approaches (AKA Start of This Week's Wonkwar on the Results Agenda," *From Poverty to Power* (blog), Oxfam, January 22, 2013, http://oxfamblogs.org/fp2p/the-political-implications-of-evidence-based-approaches-aka-start-of-this-weeks-wonkwar-on-the-results-agenda.

10 Chris Roche and Linda Kelly, "The Evaluation of Politics and the Politics of Evaluation," Developmental Leadership Program, Background Paper 11, August 2012. Chris Roche and Linda Kelly, "Monitoring and Evaluation When Politics Matter," Developmental Leadership Program, Background Paper 12, October 2012.

11 John Kania, Mark Kramer, and Patty Russell, "Strategic Philanthropy for a Complex World," *Stanford Social Innovation Review* (Summer 2014), and ongoing responses can be found at www.ssireview.org/articles/entry/strategic_philanthropy.

12 Matt Andrews and his counterparts at Harvard pioneered Problem-Driven Iterative Adaptation, which I cite throughout the paper. While my understanding of reform is more conflictual than the cooperative descriptions favored by Andrews, our conclusions regarding the most effective planning process are the same.

13 Among the most thoughtful studies assessing social and political change in the U.S. context is Steven Teles and Mark Schmitt, "The Elusive Craft of Evaluating Advocacy," white paper published by the Hewlett Foundation, July 16, 2014, www.hewlett.org/library/grantee-publication/elusive-craft-evaluating-advocacy.

14 For the full story here and of a multitude of other extractive regimes preferring to not engage in the development of roads, railroads, education, and other public goods that could threaten their hold on power, see Acemoglu and Robinson, *Why Nations Fail*.

15 Dale Russakoff, "Schooled," *New Yorker*, May 19, 2014.

16 Quoted in Bo Rothstein, *The Quality of Government* (Chicago: University of Chicago Press, 2011), 106.

17 For a dynamic chart of the resurgence of these diseases, see National Public Radio, "How Vaccine Fears Fueled the Resurgence of Preventable Diseases," January 25, 2014, www.npr.org/blogs/health/2014/01/25/265750719/how-vaccine-fears-fueled-the-resurgence-of-preventable-diseases.

18 Eric Patashnik, *Reforms at Risk* (Princeton, NJ: Princeton University Press, 2008).

19 Benjamin Friedman, in *The Moral Consequences of Economic Growth* (New York: Knopf, 2005), suggests that reforms can often gain a toehold in periods of economic growth and then face more opposition when economies contract. If true, this provides another part of the pendulum swing that program designers and evaluators could take into account.

20 In a randomized controlled trial, different groups of individuals might be divvied up and measured, with a specific intervention the main difference between the groups. While other differences might exist among individuals, once enough people are aggregated, these differences matter less and outliers can be eliminated. A rigorous regression looks to identify causation for an independent variable. To do so, the social scientist assembles a data set with multiple variables measured, and holds various variables constant or creates dummy variables to test the relative statistical significance of each potential explanatory factor.

21 Robert Jervis, *System Effects: Complexity in Political and Social Life* (Princeton: Princeton University Press, 1997).

22 See Oeindrila Dube and Suresh Naidu, "Bases, Bullets, and Ballots: The Effect of U.S. Military Aid on Political Conflict in Colombia," Center for Global Development White Paper, 2013.

23 An excellent layperson's guide to complexity theory is Melanie Mitchell, *Complexity: A Guided Tour* (Oxford: Oxford University Press, 2011), as well as Neil Johnson, *Simply Complexity* (New York: OneWorld Publishing, 2010). Ben Ramalingam and Harry Jones, with Toussaint Reba and John Young, "Exploring the Science of Complexity: Ideas and Implications for Development and Humanitarian Efforts," Overseas Development Institute Working Paper 285, October 2008, is a more useful guide to the science than Ramalingam's later book: *Aid on the Edge of Chaos: Rethinking International Cooperation in a Complex World* (Oxford: Oxford University Press, 2014). See also Patricia Patrizi and Elizabeth Heid Thompson, "Beyond the Veneer of Strategic Philanthropy," *Foundation Review* 2, issue 3 (2011); Patricia Patrizi et al., "Eyes Wide Open: Learning as Strategy Under Conditions of Complexity and Uncertainty," *Foundation Review* 5, issue 3 (2013); and Owen Barder's three-part blog (www.cgdev.org/blog/what-development) also discuss the ways in which complexity science can bring useful lessons to the development field. Last, but far from least, Vivienne O'Connor has published an excellent piece on how complexity science applies to post-conflict rule of law reform, which comes to many similar conclusions as this report in how to amend program design. See Vivienne O'Connor, "A Guide to Change and Change Management for Rule of Law Practitioners,"

Practitioner's Guide, International Network to Promote the Rule of Law, January 2015, http://inprol.org/system/files_force/publications/%5Bsite-date-yyyy%5D/inprol_pg_on_change_and_change_management_final_1_22_2015.pdf?download=1.

24 The role of the Colombian military in supporting paramilitary activity is controversial and well documented. A few of the studies include: Daron Acemoglu, James Robinson, and Rafael Santos, "The Monopoly of Violence: Evidence From Colombia," White Paper, May 2010; Oeindrila Dube and Suresh Naidu, "Bases, Bullets and Ballots: The Effect of U.S. Military Aid on Political Conflict in Colombia," White Paper, December 2013; "The Sixth Division: Military-Paramilitary Ties and U.S. Policy in Colombia," Human Rights Watch, September 2001. For the emotional state of military actors in Colombia, I draw on my own case study research, January 2015.

25 Ramalingam et al., "Exploring the Science of Complexity," note 1.

26 The parapolitics scandal in Colombia uncovered many of these illicit connections; for scholarship on the phenomenon, see Acemoglu, Robinson, and Santos, "The Monopoly of Violence."

27 See Benoit B. Mandelbrot, *The Fractal Geometry of Nature* (New York: W. H. Freeman and Company, 1982), or, for more fun, Nigel Lesmoir-Gordon, *Introducing Fractals: A Graphic Guide* (New York: Icon Books, 2005).

28 See, for instance, Denise Paquette Boots, "Neurobiological Perspectives of Brain Vulnerability in Pathways to Violence Over the Life Course," in *The Ashgate Research Companion to Biosocial Theories of Crime*, ed. Kevin M. Beaver and Anthony Walsh (Burlington, VT: Ashgate Publishing, 2011), 181–212.

29 In many ways, the international development world is now learning what the urban development world learned in the 1960s as a result of the fight between Jane Jacobs and Robert Moses in New York City. Top-down, clear, and simple urban development processes looked like the best way to develop better cities in the 1950s. The powerful Moses was their champion. But 20-story "project" buildings for the poor clustered together, highways that cut off neighborhoods from one another, and other effects of top-down, "designed" development created sterile, dead areas that concentrated poverty and exacerbated crime. Jacobs, a neighborhood activist, wrote a book articulating the messy, hidden processes and interactions that made some neighborhoods safe, functional, and successful. Her arguments eventually led to the New Urbanism movement that has revitalized cities around the world.

30 *Doing Business 2007: How to Reform* (Washington DC: World Bank, 2006), 5.

31 Zia Khan, "Response to 'Strategic Philanthropy for a Complex World,'" S*tanford Social Innovation Review* (Summer 2014), www.ssireview.org/up_for_debate/strategic_philanthropy/zia_khan.

32 Cited in Patricia Patrizi and Elizabeth Heid Thompson, "Beyond the Veneer of Strategic Philanthropy," *Foundation Review* 2, issue 3, article 6 (2011): 54.

33 Horst W. J. Rittel and Melvin M. Webber, "Dilemmas in a General Theory of Planning," *Policy Sciences* 4, Elsevier Science (1969): 155–73.

34 Margaux Hall, Nicholas Menzies, and Michael Woolcock describe such an experimentalist design and how it was used by the World Bank in their chapter: "From HiPPOs to "Best Fit" in Justice Reform: Experimentalism in Sierra Leone," in *The International Rule of Law Movement: A Crisis of Legitimacy and the Way Forward*, ed. David Marshall (Cambridge, MA: Harvard University Press, 2014).

35 Also highlighted by Roche and Kelly 2012, both papers.

36 As Daniel Kahneman writes in *Thinking Fast and Slow* (New York: Farrar, Straus, and Giroux, 2011), describing one expert-based international relations experiment in the 1990s, "People who spend their time and earn their living studying a particular topic produce poorer predictions than dart-throwing monkeys." Philip Tetlock, the author of *Expert Political Judgment* (Princeton: Princeton University Press, 2006), provides slightly higher ratings but has dedicated himself to an empirical project to determine how predictions can evolve from anecdote and "expertise" to accuracy: http://goodjudgmentproject.com.

37 Roche and Kelly, "Monitoring and Evaluation When Politics Matter."

38 Peter A. Hall of Harvard describes this beautifully in "Aligning Ontology and Methodology in Comparative Research," in *Comparative Historical Analysis in the Social Sciences*, ed. J. Mahoney and D. Rueschemeyer (Cambridge and New York: Cambridge University Press, 2003), 373–404.

39 See, for instance, Charles C. Ragin, *The Comparative Method* (Berkeley: University of California Press, 1987); Joshua Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton: Princeton University Press, 2008); David Freedman, "Statistical Models and Shoe Leather," in *Sociological Methodology*, vol. 21, ed. Peter V. Marsden (Washington DC: American Sociological Association, 1991); and David Freedman, "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review* 30 (2006): 691–713.

40 Peter A. Hall, "Aligning Ontology and Methodology in Comparative Politics," in Mahoney and Rueschemeyer, eds., *Comparative Historical Analysis in the Social Sciences*, 373–406, 385.

41 See Adnan Khan, Asim Khwaja, and Ben Olken, "Property Tax Experiment in Punjab, Pakistan: Testing the Role of Wages, Incentives, and Audit on Tax Inspectors' Behavior," Poverty Action Lab, www.povertyactionlab.org/evaluation/property-tax-experiment-punjab-pakistan-testing-role-wages-incentives-and-audit-tax-inspe.

42 See Poverty Action Lab, "Political Economy and Governance," www.povertyactionlab.org/political-economy-governance. The document details RCTs done on political issues, and suggests that the approach works best when the government already has the "will"—while doing little to unpack that complex term or demonstrate ways of affecting government will. One thing the Poverty Action Lab's findings suggest is that implementation by bureaucracies is less generalizable than other RCT findings. A handful of RCTs and natural experiments are beginning to focus on the "how" of implementation. Though it is unclear how generalizable these findings are, they are a step in the right direction. See, for instance, C. Ferraz and F. Finan, "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes," *Quarterly Journal of Economics* 123, no. 2 (2008): 703–45, who look at the effect of public audits on electoral accountability for corruption in Brazil; A. Banerjee, S. Kumar, R. Pande, and F. Su, "Do Informed Voters Make Better Choices? Experimental

Evidence From Urban India," 2010, www.snsindia.org/jpal_impact_evaluation.pdf, who consider the role of disclosing information about incumbent performance in reelection in Delhi and in Uttar Pradesh; and Alberto Chong, Ana L. De La O, Dean Karlan, and Leonard Wantchkon, "Looking Beyond the Incumbent: The Effects of Exposing Corruption on Electoral Outcomes," 2013, who consider providing information about corruption on electoral outcomes in Mexico. The mixed findings show some of the difficulties of RCTs in process-tracing reform implementation and cross-study comparison, though they remain an excellent start toward accumulating a body of knowledge about a particular type of intervention.

Other excellent RCT studies show the difficulties with time-delineation of such studies. For instance, A. Banerjee, E. Duflo, and R. Glennerster, "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System," *Journal of the European Economic Association* 6, issue 2–3 (2008): 487–500, shows that while monitoring and financial incentives positively impacted health performance in rural India for its first six months, the finding altered in the subsequent six months, likely because the local health administration took steps to deliberately undermine the incentive system. In another case, the Poverty Action Lab finds that "community participation" has mixed results. This is hardly surprising to anyone who has engaged in community participation work—there is a world of difference in programming between tokenistic community consulting and full community leadership. Moreover, communities differ—for instance, in Honduras's Mosquito Coast, where narcotraffickers have gained control of most local NGOs and the leading indigenous community organization, community consultation with major stakeholders is likely to present the opinions of narcotraffickers through their hand-selected representatives. Such findings confirm Matt Andrews's critique that experimental results "present sanitized outcomes, whose replication when scaled-up under less-sanitized, real-world conditions are questionable." Matt Andrews, "The Richest 30 Countries in the World in 2040 Are…" *Governance Reform in International Development* (blog), http://matthewandrews.typepad.com.

43 Angus Deaton, "Instruments of Development: Randomization in the Tropics, and the Search for Elusive Keys to Economic Development," Keynes Lecture, British Academy, 2009. James J. Heckman and Sergio Urzua, "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, 156:1 (May 1, 2010): 27–28, suggests that these approaches often gain precision by asking narrower questions and are therefore "less ambitious in the range of questions they seek to answer"—a concern shared by Francis Fukuyama in "Political Order in Egypt," *American Interest* (May/June 2011).

44 See, for instance, Seth D. Kaplan, *Betrayed: Politics, Power, and Prosperity* (New York: Palgrave, 2013).

45 *Fighting Corruption in Public Services: Chronicling Georgia's Reforms* (Washington DC: World Bank, 2012), http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2012/01/20/000356161_20120120010932/Rendered/PDF/664490PUB0EPI00657 74B09780821394755.pdf.

46 "Arbitrary Removal of Guatemala Attorney General Claudia Paz y Paz," press release, Robert F. Kennedy Center for Justice & Human Rights, February 18, 2014, http://rfkcenter.org/arbitrary-removal-of-guatemala-attorney-general-claudia-paz-y-paz-2.

47 Frances Z. Brown, "Rethinking Afghan Local Governance Aid After Transition," United States Institute of Peace Special Report, August 2014, 9, www.usip.org/sites/default/files/SR349_Rethinking-Afghan-Local-Governance-Aid-After-Transition.pdf.

48 See Roche and Kelly, "The Evaluation of Politics and the Politics of Evaluation."

49 See USAID, *Macedonian Business Resource Center (MBRC), Republic of Macedonia, Final Report STO no. 3*, January 2001–December 2002, http://pdf.usaid.gov/pdf_docs/Pdack644.pdf.

50 See the "Performance Evaluation of the USAID/Vietnam Support for Trade Acceleration (STAR) Project, Final Report," May 2011, 11, http://pdf.usaid.gov/pdf_docs/pdacs486.pdf: "The GVN's [government of Vietnam's] Law on Laws is an example of a law that went beyond addressing individual commercial laws to altering the foundation of the legal system. Several interview respondents in the legal profession noted that the Law on Laws was the GVN's single most important legislative accomplishment. A major component of the Law on Laws was the requirement for publication of drafts with concomitant opportunity for public comment. Regular adherence to this provision, with the assistance of STAR, was instrumental in changing the legislative culture. Another critical initiative pertains to codification of laws. STAR II was instrumental in the GVN including the codification provision in the Law on Laws. An interview respondent noted that: In the last five years there has been unambiguous improvement in making the law process transparent and participatory. There has been a definite improvement. STAR has been very active in this area."

51 Roche and Kelly, "The Evaluation of Politics and the Politics of Evaluation."

CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE

The **Carnegie Endowment for International Peace** is a unique global network of policy research centers in Russia, China, Europe, the Middle East, and the United States. Our mission, dating back more than a century, is to advance the cause of peace through analysis and development of fresh policy ideas and direct engagement and collaboration with decisionmakers in government, business, and civil society. Working together, our centers bring the inestimable benefit of multiple national viewpoints to bilateral, regional, and global issues.

The **Carnegie Democracy and Rule of Law Program** rigorously examines the global state of democracy and the rule of law and international efforts to support their advance.

BEIJING BEIRUT BRUSSELS MOSCOW WASHINGTON

THE GLOBAL THINK TANK

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org