CMI CHR. MICHELSEN INSTITUTE

U4 ISSUE

August 2013 No 8

Methods for learning what works and why in anti-corruption

An introduction to evaluation methods for practitioners

Jesper Johnsøn
Tina Søreide

U4 Anti-Corruption Resource Centre
www.U4.no

U4 is a web-based resource centre for development practitioners who wish to effectively address corruption challenges in their work.

U4 is operated by the Chr. Michelsen Institute (CMI) – an independent centre for research on international development and policy – and is funded by AusAID (Australia), BTC (Belgium), CIDA (Canada), DFID (UK), GIZ (Germany), Norad (Norway), Sida (Sweden) and The Ministry of Foreign Affairs Finland.

Methods for learning what works and why in anti-corruption

An introduction to evaluation methods for practitioners

Jesper Johnsøn
Tina Søreide

U4 Issue
August 2013 No 8

Contents

Abstract

Evaluations of donor-funded anti-corruption reforms and programmes would benefit from upgrading and diversifying the methods used to document effects. Better evaluations in turn would improve the evidence base for the effectiveness of specific anti-corruption interventions. Using real and hypothetical examples, this paper offers practical guidance to practitioners who design, implement, and disseminate evaluations and research on anti-corruption. A range of quantitative and qualitative methods can be used to answer operational questions on the impact of anti-corruption interventions. Some methods can produce stronger evidence than others for a specific evaluation, but there are trade-offs between rigour and costs, and between aspiration and feasibility. Donors should let the evaluation question, programme attributes, and availability of data determine the most appropriate methods for a given study. With careful planning and adequate resources, donors can use many of the methods presented in this paper. This should give more reliable results and produce needed knowledge on what works in anti-corruption, laying the basis for more effective anti-corruption initiatives.

Acknowledgements

This issue paper addresses policy challenges frequently discussed in the anti-corruption community. We are grateful that our understanding of these challenges could develop in dialogue with numerous anti-corruption experts and programme managers internationally. In particular, we want to thank the U4 partners who made this study possible, as well as Geir Sundet and Hannes Hechler, for their formative inputs and comments at an early stage of the document. We also thank the reviewers of this study, Eirik Gjøstein Jansen, Ida Lindkvist, Linda Payne, Vincent Somville, and Charlotte Vaillant, who provided excellent advice. Our colleagues at CMI and U4 contributed useful comments, especially Liz Hart and Ivar Kolstad. Any remaining weaknesses in the paper are our own responsibility.

Abbreviations

CRC	citizen report card
CSO	civil society organization
ICT	information and communications technology
OECD/DAC	Development Assistance Committee of the Organisation for Economic Co-operation and Development
PETS	Public Expenditure Tracking Survey
QSDS	Quantitative Service Delivery Surveys

1. Introduction

Governments, development partners, and researchers have long been hindered by a lack of evidence about the impact of specific anti-corruption interventions. They have ample anecdotal information about what works, but rarely are they able to prove which good governance and anti-corruption initiative was decisive in bringing about change. In order to develop our knowledge of the effectiveness of different approaches and defend funding of governance and anti-corruption initiatives, we need to be able to isolate their impact.

This paper offers practical guidance to practitioners who design, implement, and disseminate evaluations and research on anti-corruption, as well as to those who oversee support to anti-corruption programmes or advocate for funding. Drawing on real and hypothetical examples, it shows a range of methods that can be used to answer operational questions regarding the impact of an anti-corruption intervention and suggests how evaluators, programme managers, and donor staff can use these methods.

Corruption is an area where results can be difficult to define and even more difficult to achieve. For task managers, a loosely formulated goal will make failure less evident. In this sense, aiming for "increased awareness" or "more transparency" has been safer than attempting to define specific outcome-level objectives such as less fraud, fewer bribes, more convictions, better institutional integrity scores, and so on. Over the past decade, nevertheless, donor agencies have become much better at defining goals and determining indicators of success (Liverani and Lundgren 2007, 241–55). We hope this paper will strengthen that trend in the anti-corruption sector.

Since the fight against corruption started to be taken more seriously a couple of decades ago, a substantial amount of empirical work and research has been undertaken. The harmful effects of corruption on development are well documented (see, for example, Svensson 2005). Research, often financed by development partners, has provided guidance for practitioners as to which countries, sectors, and circumstances present the greatest corruption risks. Wide civil servant discretion, asymmetric information, market and governance failures, and access to external rents—such as revenues from extractive industries or foreign aid—have been identified as important facilitating factors for corruption (Campos and Pradhan 2007; Johnston 2005; Rose-Ackerman 1978, 1999; Klitgaard 1988). Corruption issues can persist due to various failures in control and integrity systems, including lack of checks and balances at the political level, collective action challenges, weak law enforcement, low trust in government institutions, and weaknesses in state legitimacy, the latter reflected in patronage networks, loyalty to ethnic groups rather than institutions, and democratic failure (Clapham, 1985, Pope 2000; Rothstein 2011).

Controls and sanctions do reduce the individual inclination to become involved in corruption. Moreover, access to information, competition in business and politics, and safe channels for whistleblowing have been shown to discourage corruption, given the right preconditions (Rose-Ackerman and Truex 2012; Hunt 2006; Treisman 2000; Lambsdorff 2006).

While knowledge about corruption, generated by researchers and documented by practitioners, has increased substantially, producing *evidence* that anti-corruption interventions have an impact in reducing corruption is a relatively new area for research and evaluation. Some empirical academic research has used methods that produce strong evidence of impact and are operationally relevant, such as work by Björkman and Svensson (2009), Olken (2007), and Reinikka and Svensson (2003). However, researchers do not always seek to answer the operational questions posed by aid donors, and a handful of good studies does not produce sufficient guidance for anti-corruption initiatives in general (see Ravaillon 2009b). Anti-corruption practitioners are still searching for answers on how best to translate principles such as sanctions, control, transparency, and accountability into reforms

and programmes that can reduce corruption in public service delivery and political systems. For example, the question of whether formal audit or community monitoring is the best control mechanism for public service delivery is still controversial, and existing research provides mixed messages (Hanna et al. 2011, 4–5; Johnsøn, Taxell, and Zaum 2012, 43). Debate also continues on whether isolated anti-corruption initiatives in a public service reform setting can work when they are not part of a larger reform process or coinciding with economic growth (Khan, 2012; Baj et al., 2013)

These operational questions are the new frontier for anti-corruption research. Donors and researchers have to reformulate the questions they pose, as well as the methods they use to generate knowledge. These groups need to work together, rather than in parallel, by integrating programme design with research design through operational research. Equally important, in order to obtain answers to operational questions, evaluators should play a greater role in generating knowledge on what works and why—a far more ambitious task than the quick ex-post evaluation typically commissioned by donors.

This paper shows how good evaluation methods can be used at the project level in a manner that is feasible for practitioners and fits within a reasonable price range. Given the vast universe of anti-corruption interventions and reforms, donors need a *cost-effective strategy* for producing the best possible evidence base. Such a strategy would rely on different types of evaluation methods at different costs. Roughly categorized, there are two main groups of evaluations: on the one hand, a very small group of rigorous and highly publicized experimental/quasi-experimental studies, and on the other hand, a very large group of evaluations conducted under budget and time constraints with weak methodologies (Bamberger 2009, 2–3). To overcome this divide and the "evaluation gap," this paper presents a menu of evaluation options suited to different levels of resources and data availability.[1]

Evaluations of anti-corruption interventions can benefit from improvements in both methodological design and methods. This paper focuses exclusively on evaluation *methods*, a variety of which are easily available to practitioners.[2]

By introducing various possible evaluation methods for anti-corruption activities, we wish to highlight the diversity of alternative approaches and encourage creative solutions to challenges in evaluating anti-corruption reforms. At the same time, we reaffirm the social science doctrine that the choice of evaluation design should depend on the nature of the intervention being studied and the questions one wishes to answer. In short, a combination of three factors will determine which design and methods are most useful: (a) the nature of the evaluation question, (b) programme attributes and how people are exposed to the programme or policy, and (c) the available data. Sometimes these factors will call for either quantitative or qualitative methods, but in most cases a mixed-methods approach will yield

[1] Savedoff, Levine and Birdsall (2006) define this gap as "the missing body of impact evaluation work that is required to guide future social development policy."

[2] This distinction between design and method is not so tidy in practice. Real-world designs are almost always hybrids, rarely pure types. Of special importance for this paper is the fact that an experiment is a design, but randomization is a data collection technique. Our paper focuses on randomization. We use the definitions of methodological design and methods provided by Elliot Stern and colleagues. *Design* is defined as the overarching logic that informs how an evaluation is conducted, consisting of four elements: evaluation questions, theory used to analyse data, data, and use of data. Evaluation designs are commonly classified as experimental, quasi-experimental, or nonexperimental. but theory-based evaluation and case studies also qualify as types of design. *Methods* are approaches to data collection, measurement tools, and statistical analysis (Stern et al. 2012, 15).

the most reliable results. The range of data collection methods that exist can be better exploited for evaluations of anti-corruption work.

The paper loosely categorizes guidance on different methods (not designs) into *gold*, *silver*, and *bronze*. The categories refer to the methods' potential to provide reliable data on what works. This relates to the concept of *validity* of findings. An evaluation can be assessed in terms of its internal validity (are effects correctly estimated for the specific case under study?) and its external validity (can we generalize from the specific case and assume that similar results will occur in other contexts?). Gold methods have the potential to produce evaluations with stronger internal and external validity than silver, and silver methods can produce stronger data than bronze, with the significant caveat that the research or evaluation must be correctly implemented and free of bias, have adequate sample sizes, and so forth. Gold is not the only method of value to policymakers; it neither applies to all contexts nor answers all questions that policymakers want answered. The aim of donor agencies should be to increase the total tally of evaluations using gold, silver, and bronze methods, preferably combined, to strengthen the overall evidence base.

The structure of the paper is as follows:

- Section 2 presents ways to improve foundations for evaluations and deal with complex activities.
- Section 3 explains three basic types of evaluation—impact, programme, and process—and their purposes, basic principles, and corresponding methods.
- Section 4 shows how randomization and field experiments can be used to show the impact of anti-corruption interventions of any size. Such methods are sometimes referred to as the gold standard. The section also addresses the main limitations of this method.
- Section 5 shows how to use statistical matching methods and advanced surveys to produce rigorous evidence. These methods are collectively labelled silver.
- Section 6 addresses the problems facing the large majority of evaluations in anti-corruption: how to evaluate programme effects when one has insufficient data, time, and/or budget to apply the silver or gold methods. Bronze methods can increase the overall quality of an evaluation of an ongoing or finished programme.
- Section 7 outlines principles for strengthening evaluation designs for anti-corruption interventions by using mixed methods and triangulation.

2. Laying the groundwork for good evaluations

The first two steps in any evaluation of an anti-corruption intervention are to:

- Think through how the initiative is intended to contribute to behavioural, organizational, political, or societal change, and express this process clearly, for example, in a *results chain* or *theory of change*; and
- Work on increasing the *evaluability* of the reform or programme, for example, by establishing indicators and arranging for data to be collected systematically and consistently.[3]

Reflection on these questions will improve the quality of evidence across the gold, silver, and bronze methods explained below. In this reasoning process, it is often helpful to break down complex interventions into their constituent parts and to make use of formative evaluation processes. Besides strengthening the evaluations of reforms, such groundwork also generally enhances the quality of policies and programmes by testing their internal logic.

2.1 Increasing evaluability, measuring change, and understanding impact

Programme managers should be prepared to integrate evaluation considerations into the overall design of an anti-corruption programme. Ensuring that evaluations become an integral part of the initiative is essential for a good result. The most effective methods of evaluating anti-corruption interventions are those where monitoring mechanisms are built into the projects from the beginning.

Instruments for data collection and dissemination are often a key component of successful anti-corruption interventions. Examples include citizen report cards and other systems of community monitoring such as those documented by Olken (2007) in Indonesia. Some basic preconditions for evaluating an intervention include having clear goals/objectives, measurable performance indicators, a baseline, and an operational monitoring and evaluation system able to regularly collect data on indicators. For example, just the creation of well-designed indicators and collection of baseline data substantially increases the potential for producing valuable knowledge and evidence. Various manuals offer guidance on how to apply evaluation quality standards, including one produced by the Development Assistance Committee of the Organisation for Economic Co-operation and Development (OECD/DAC 2010b). Only a few manuals have been produced specifically for use in anti-corruption, with Johnsøn et al. (2011) as one example.

To lay the groundwork for evaluation of an anti-corruption initiative, planners must consider in detail how change will happen and identify exactly which direct and indirect changes the anti-corruption initiative is expected to trigger. The preparatory work should identify the intended sequence of steps towards anti-corruption impact—that is, the results chain—and the indicators that will be used to measure the effects at each step. Analysing the results chain and indicators for a given intervention will often bring underlying assumptions into focus.

[3] *Evaluability* refers to how well an intervention (policy, reform, programme, project) can be evaluated. Establishment of baselines, comparison groups, systematic data collection, and a clear programme logic can all help improve evaluability.

In the following two examples, a results chain has been drawn for a hypothetical anti-corruption intervention, with indicators specified for each part of the chain. However, one example shows a relatively weak and imprecise results chain, and the other a chain that is stronger.

Example 1: Grant facility for civil society organizations to increase accountability in public service provision

In this example, shown in table 1, a grant facility for civil society organizations is established with the intended aim of promoting better service delivery and reducing corruption. Table 1 illustrates the assumptions underlying each step.[4]

The first column presents the specific problem that motivates an intervention—in this case, weak service delivery that is hampered by corruption. Input is what the donor-financed project provides: in this example, it is financial support to civil society organizations (CSOs) so they can conduct advocacy campaigns that promote pro-poor reforms. The output *might* be more CSOs with more competent staff who do in fact conduct such campaigns, but that effect depends on how the CSOs spend the funds from the donors. Moreover, one can then hope, but not take for granted, that the advocacy campaigns result in the greater political accountability through elections and that this increased accountability in turn leads to reduced corruption in service delivery. But all of these causes and effects are highly uncertain. What we find, by placing each step in a table like this, is that the expected impact of providing financial support to CSOs is based on a chain of loose assumptions that can be difficult to validate using indicators. Even if positive results appear, it would be difficult to attribute these changes to the grant facility with certainty.

Table 1: Grant facility for civil society organisations: Results chain

Problem	Input	Output	Outcome	Impact	Long-term goal
Weak service delivery (health, education, utilities) due to corruption and poor political accountability	Financial support to CSOs	More CSOs with more qualified governance staff Advocacy campaigns	Democratic elections result in more seats for accountable politicians	Accountable and reform-friendly politicians Good governance	Better service delivery, less corruption
Indicators					
Corruption cases and news stories Discrepancy between allocation and actual use of funds	Number of CSOs Number of newspaper articles, TV interviews, anecdotal success stories, etc.		Public trust in politicians (Global Corruption Barometer, Afrobarometer, etc.) Economist Intelligence Unit Democracy Index scores		Millennium Development Goal indicators

Source: Adapted from Abdul Latif Jameel Poverty Action Lab.

[4] While the specific example is our idea, the table is derived from materials presented at J-PAL Europe, a workshop on evaluating social programmes organized by the Abdul Latif Jameel Poverty Action Lab (http://www.povertyactionlab.org/) in Brussels, 12–16 September, 2011, and attended by CMI staff member Tina Søreide.

Example 2: Rewarding whistleblowers to reduce corruption in customs

Consider now a different case: an anti-corruption programme designed to reduce corruption in customs by offering a reward to those who report that they have paid a bribe. There should be no sanction for those who report the bribery.[5] Those who have demanded a bribe will have to pay a fine that is significantly greater than the amount of the bribe. Reading the results chain in table 2, we realize that in this case we have a much clearer idea of how an impact evaluation could be conducted.

Table 2: Rewarding whistleblowers to reduce corruption: Results chain

Problem	Input	Output	Outcome	Impact	Long-term goal
Firms complain about costs due to corruption in customs	Information provided to firms that they will be awarded the amount of their bribe if they report the crime	More firms report demands for bribes in customs	Customs officials sanctioned more often for demanding bribes	Less bribery in customs	More business-friendly and competitive investment climate
Indicators					
Corruption cases and news stories	Whistleblower policy approved Information disseminated Safe reporting mechanism established	Number of reported cases	Number of sanctioned customs officials	Number of reported cases Average estimate of bribery costs by firms	Scores on World Bank's Doing Business or Enterprise Surveys

Source: Adapted from Basu 2011.

These examples of an CSO grant facility and a whistleblowers programme used results chains to clarify the logic of the interventions—how one change leads to the next—and identify indicators at each step. These indicators are decisive in documenting causality: in the first example, how advocacy can improve service delivery and lower corruption, and in the second example, the extent to which reduced corruption in customs can be attributed to the whistleblower programme.

We find that a credible impact evaluation is possible in the second example and not in the first. It turns out that the grant facility programme is based on a much greater leap of faith than the customs programme, as success is preconditioned on weakly specified democratic processes, political will, and the performance of government agencies responsible for service delivery. It seems overambitious to expect such a grant facility by itself to positively influence all these actors and factors. For such a programme to receive donor funding, it should be able to more credibly explain its theory of change.

The practice of developing results chains like those illustrated above would provide an improved foundation for many anti-corruption evaluations because it "forces" the task manager to think through

[5] All players' incentives must be understood in light of the given institutional context. It is important to structure the programme and reward scheme to that bribe-givers do not have incentive to over-report. The reward could therefore be a fixed amount based on the number of complaints and average bribe payments.

underlying assumptions behind each step towards change. An even stronger foundation would come from constructing a *theory of change*, although this method is slightly more costly in time and resources. The latter approach has been developed specifically for evaluation of complex interventions and is considered to be a useful tool for the governance and anti-corruption sector (White 2009). In recent meta-evaluations, Norad and the World Bank have criticized donor-financed anti-corruption interventions for lacking a coherent theory of change (Norad 2011; IEG 2011). A main message in this criticism is that without clear, explicit logical frameworks for how an initiative is supposed to bring change, the meta-evaluations could not bring much new information about what works and what does not.

Theory of change goes beyond the logframe approach by not only considering how the interventions' inputs and results are linked, but also analysing the causal chain, preconditions, and assumptions required for change to happen (Funnell and Rogers 2011, xix–xx, 22). Hence, compared to the results chains illustrated above, a theory of change implies a more precise specification of the preconditions necessary for each link in the causal chain to work, as well as a detailed understanding of critical underlying assumptions.[6] It also includes considerations of the socioeconomic and political-economic context for the intervention.

The main point is that anti-corruption implementers and evaluators should have a road map for how to recognize change. Drawing a results chain and developing a theory of change are, to different degrees, useful in understanding the expected changes and preparing a theory-based evaluation approach. This further enables donors to claim causality.

2.2 Dealing with complexity and the small *n*

Governance and anti-corruption interventions are rarely easy to evaluate. Their complex character and the fact that many interventions have only one or a few target units makes comparisons and statistical tests difficult. Three strategies can help overcome such complexity.[7] First, as explained in section 2.1, using theory-based evaluation tools can improve evaluability by making the programme logic explicit. Two other strategies are discussed in this section: (a) breaking down complex interventions into components and (b) using *formative* evaluation. The latter implies building feedback loops into programme design and learning throughout an initiative, in addition to assessing the end result.

A lack of comparable data makes it difficult to conduct solid evaluation. White and Phillips (2012) consider how to address attribution of cause and effect in cases where a reform targets only one or a few units, such as individuals or organizations. An example could be the effects of reform of a civil service commission. In such cases, comparisons are difficult and statistical analysis is often irrelevant. The authors describe a range of approaches which can be used in such "small *n*" cases. These use a different basis for causal inference than do experimental or quasi-experimental approaches. Rather than using a statistical counterfactual, such approaches attribute changes to the intervention by examining the underlying processes and mechanisms that can explain cause and effect, testing for rival explanations if possible. The establishment of a theory of change or other programme logic is key to these approaches. As formulated by White and Phillips: "Causation is established by collecting evidence to validate, invalidate or revise the hypothesised explanations, with the ultimate goal of rigorously documenting the links in the actual causal chain" (2012, 28).

[6] For more information about theory of change analysis, including detailed approaches and case studies, see Johnsøn (2012).

[7] For more advanced literature on complexity, see Barder (2012), Pritchett, Woolcock, and Andrews (2010), and Andrews (2013).

One possible reason why the evidence base for anti-corruption reforms is thin is that most evaluations so far have focused on proving the impact of either specific institutions (for example, anti-corruption authorities) or complex policies (typically anti-corruption policies). These interventions are normally overarching frameworks encompassing a variety of activities and goals relating to anti-corruption. Evaluators thus find it difficult to go beyond general statements of performance. One way to deal with complexity is to break down the overall intervention into its constituent parts. For example, when evaluating an anti-corruption agency, it can be useful to first evaluate separately individual work streams such as investigation, prosecution, public awareness, and corruption prevention activities.[8] Different evaluation methods can be used for each of these diverse activities.

In seeking to understand what works and why in anti-corruption, it might be tempting to focus on larger reforms or initiatives, for example whether anti-corruption agencies in general have a positive or negative effect. However, it is even more useful to assess the effectiveness of a certain set of activities—for example, an agency's public awareness activities—in reducing corruption. If we can choose, it is often better to assess several specific mechanisms and not only the overall performance. We need to isolate activities in order to isolate and attribute effects. In the health sector, for example, organizational assessments of the Ministry of Health or individual hospitals do not automatically provide proof that specific interventions are working or not. Similarly, for anti-corruption we need organizational assessments but also studies of specific interventions and their effectiveness. Donors typically fund programmes or institutions, but the effectiveness of that support will in the end depend on the effectiveness of individual projects or activities, such as service delivery, community monitoring, process reengineering, and public awareness campaigns. At this level, opportunities for documenting behavioural change are greater.

Another way to deal with complexity is to grant the evaluator a larger role in the design and implementation of the process through the use of formative evaluation, sometimes called real-time evaluation.[9] Evaluations have two important purposes: Providing results about what works (*learning*) and checking if implementers are doing their jobs (*accountability*). Often accountability aspects are more prominent than those centred on learning outcomes. This relates to the difference between formative and summative evaluations.

Most evaluations of development interventions are *summative* evaluations, assessing the eventual effect of a programme and whether it was the programme that caused the desired outcomes/impacts. The other type of evaluation, *formative* evaluation, is hardly used in the area of anti-corruption. Formative evaluation helps improve evaluability and the actual design of the intervention as it is being implemented. It takes an active part in forming the programme by examining the delivery, the implementation process, and the enabling/constraining environment.[10] Of the three types of evaluations mentioned in section 3, process evaluation is a formative evaluation, while outcome and impact evaluations are typically only done as summative evaluations. The principles of good formative evaluation, however, go beyond the typical process evaluation and focus more on pre-testing of the programme, its evaluability, and the existence of a robust monitoring and evaluation

[8] See Johnsøn (2012) for an analysis using the Malawi Anti-Corruption Bureau as an example. See also Ravallion (2009a) and Johnsøn et al. (2011) for general thoughts about evaluating anti-corruption authorities.

[9] Definitions of what constitutes a real-time evaluation vary, but most consider the distinctive features to be that the evaluator is part of an ongoing process of reflection (Norad 2010, 9), or that the "primary objective is to provide feedback in a participatory way in real time (i.e. during the evaluation fieldwork)" (Cosgrave, Ramalingam, and Beck 2009, 10).

[10] For a discussion of real-time and prospective evaluation in general, see IEG (2011).

framework with relevant indicators. This often requires that the evaluator be part of the programme design team. By contrast, the typical process evaluation is normally a form of midterm review.

3. Different types and purposes of evaluations

This section presents three main types of evaluation—impact evaluation, programme evaluation, and process evaluation—which differ in purpose. *Impact evaluation* assesses the causal effects of a programme, measuring what net change can be attributed to it (Rossi, Lipsey, and Freeman 2004, 54; Gertler et al. 2011, 7–8). *Programme evaluation* assesses whether the programme has achieved its objectives and the effectiveness and efficiency with which it has pursued these objectives. *Process evaluation* investigates how the programme is being implemented and whether its activities are performed according to plan.[11]

3.1 Impact evaluations

Impact evaluations seek to answer cause-and-effect questions. They analyse effects that are directly attributable to the implemented programme, asking whether one can claim that a given change is caused by the programme. The focus on causality and attribution determines the methods that can be used and requires estimation of a *counterfactual* (Gertler et al. 2011, 7–8).

The approach usually used is simple in principle: we compare observed performance after programme implementation with what we think *would have happened* without the programme—that is, with the counterfactual. The difference between these scenarios—one real, the other speculative—is the estimated net effect of the programme. The strength of the evidence depends on whether it is possible to compare similar groups or units that have been exposed and not exposed to the anti-corruption intervention.

As explained in box 1, difficulties arise because perfect comparison of what happens with and without the intervention requires that the two cases be identical in all ways other than exposure to the intervention. In reality, creating a perfect counterfactual scenario is seldom possible. A central challenge for the evaluation design is thus to identify the best possible counterfactual, even if it has some weaknesses. This paper's references to gold and silver reflect the validity and reliability of such comparisons.[12]

Anti-corruption interventions often seek impacts that are difficult to observe and measure: for example, improved public integrity is a less quantifiable measure than improved health. Moreover, target units in anti-corruption often are not easily compared. For example, it is harder to compare the impact of anti-corruption reforms on a ministry than it is to compare the impact of a health intervention on individual health, as there are many people with similar characteristics who can be compared but few ministries that are the same.

Very few studies of anti-corruption interventions are based on such approaches to impact assessment (see section 4). To our knowledge, no impact evaluations of donor-funded anti-corruption programmes have yet been done. Possible explanations might be a widespread belief that corruption cannot be measured or perhaps a resistance to ensuring a strong evaluation framework for anti-corruption initiatives, given the challenges in demonstrating a positive impact.[13] However, an

[11] Rossi, Lipsey, and Freeman (2004) also present other kinds of assessments, such as assessment of need and assessment of programme design and logic/theory, which are not covered in this section.

[12] Methods within both categories can use comparison groups to create counterfactual scenarios, but some methods are more precise than others. Comparison groups can be created using (a) randomization, (b) pipeline approach, (c) matching areas on observables, and (d) propensity score matching. White (2006, 12-13)

[13] In his article on the political economy of evaluations, Pritchett (2002) hypothesizes that donors resist doing scientifically rigorous impact evaluations because of the perceived high risk that the intervention will not be seen as successful. The reputational risk of a negative evaluation is seen to outweigh the reputational risk of not having a rigorous system of evaluation in place.

important distinction should be made between our ability to measure corruption levels and the feasibility of evaluating the impact of anti-corruption projects. While corruption, as a complex, largely hidden social phenomenon, will always remain hard to measure, measuring impact is possible once we have defined a counterfactual scenario and appropriate performance indicators. For this, several gold and silver methods are available, as shown in sections 4 and 5. Regarding measurement of corruption, several authors have shown that precise measurements are possible (Kaufmann 1997; Reinikka and Svensson 2006; Olken and Pande 2011) and that a useful strategy for dealing with complexity is to triangulate indicators to increase construct validity (Johnsøn et al. 2011, 41).[14]

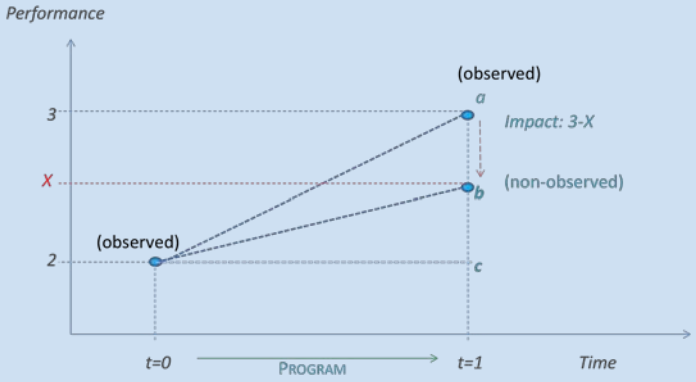
Impact questions are typically answered using experimental evaluation designs. Randomization and statistical analysis are the *methods* most often used to establish control groups and thus to create a counterfactual scenario. However, counterfactual scenarios can also be established by nonstatistical means, and causal questions can be answered using nonexperimental designs. Such methods are, however, considered "alternative" and in need of additional methodological testing (Stern et al. 2012, 7–8).

BOX 1: ESTIMATING THE COUNTERFACTUAL: WHAT WOULD HAVE HAPPENED IN THE ABSENCE OF THE PROGRAMME?

For reliable impact evaluation, one can measure the difference between what *happened* with the programme in place (ACTUAL) and what *would have happened* without the programme (COUNTERFACTUAL) to assess the net effects of the intervention (IMPACT). Displayed as an equation:

ACTUAL – COUNTERFACTUAL = IMPACT

Consider the figure below, with time and programme performance on the axes. After programme implementation, performance is estimated at point *a* (performance = 3). If improved performance were due to the programme only, the impact would be *a–c* (performance = 3–2). However, we do not know what would have happened without the programme. Possibly there would have been some improvement in performance *even without* the programme. In order to estimate performance, therefore, we have to compare the end result with what we think would have happened in the absence of the intervention. This is the counterfactual, illustrated by point *b* (performance = an unknown *X*). The challenge is to make a qualified estimate about this *X* and then make the comparison with observed performance after programme implementation.



Source: J-PAL Europe, Abdul Latif Jameel Poverty Action Lab, Brussels, 12–16 September, 2011.

[14] Construct validity refers to the extent to which measures adequately reflect what is supposed to be evaluated. Corruption has multiple manifestations, so multiple measures are necessary to ensure high construct validity, unless one focuses on a specific type of corruption such as bribery.

3.2 Programme evaluations

With a view to raising the standards and norms for programme evaluations, the Organisation for Economic Co-operation and Development through its Development Assistance Committee (OECD/DAC) has formulated five evaluation criteria: relevance, effectiveness, efficiency, impact, and sustainability.[15] These can all be covered in a programme evaluation, but questions around impact, and to some extent sustainability, usually require the use of impact evaluation methods, using statistical analysis of data collected over time. Questions regarding programme relevance, effectiveness, and efficiency can, however, be answered without using such methods when large, longitudinal data sets are not available.

While impact evaluations focus on the impact an intervention has had on long-term socioeconomic objectives, such as reducing corruption levels, most programme evaluations aim to produce evidence of whether the programme has achieved its stated objectives.[16] An impact evaluation necessitates constructing a counterfactual scenario, usually by means of comparison groups, whereas programme evaluations assess whether the intended objectives have been reached without necessarily considering what would have happened in the absence of the intervention.

Programme evaluations are sometimes called outcome evaluations, since they typically focus on programme outcomes rather than impacts. Evaluation questions often relate to issues of efficiency and effectiveness (at the output and outcome levels) and do not consider whether outcomes lead to long-term, sustainable effects (impacts). For efficiency and effectiveness assessments, comparisons are often useful. However, the comparisons are different from the ones done in impact evaluations. Programme evaluations can benefit from comparing, or benchmarking, similar units—for example, the number of prosecutions per investigator for an anti-corruption authority in country X compared with the number in country Y—without necessarily establishing a counterfactual scenario.

While programme outputs can be directly observed, measured, and documented by the programme's own monitoring and evaluation system, any conclusions on outcomes will normally require additional data collection through sample surveys, focus group discussions, and so forth (Bamberger, Rugh, and Mabry 2006, 40).

[15] Following the OECD/DAC definition, *relevance* refers to the extent to which the objectives of a development intervention are consistent with beneficiaries' requirements, country needs, global priorities, and partners' and donors' policies; *effectiveness* means the extent to which the development intervention's objectives were achieved, or are expected to be achieved, taking into account their relative importance; *efficiency* is a measure of how economically resources and inputs (funds, expertise, time, etc.) are converted to results; *impact* is defined as long-term effects (positive and negative, primary and secondary) produced by a development intervention, directly or indirectly, whether intended or unintended; and *sustainability* is an expression of the continuation of benefits from a development intervention after major development assistance has ended, the probability of continued long-term benefits, and/or the resilience to risk of the net benefit flows over time (OECD/DAC 2010a, 20–24, 32, 36).

[16] In this paper we distinguish between project and programme evaluations only when relevant for a particular argument. Evaluation designs will differ according to the complexity, scale, and scope of the programme, and of course the specific objective of the evaluation.

3.3 Process evaluations

Process evaluations are assessments of ongoing activities and of the outputs they produce. A process evaluation is commonly done as a short-term, summative exercise; arguably, the vast majority of donor-funded evaluations of anti-corruption activities fall within this category. A version popular with donors is the midterm review. Unfortunately, such exercises often emphasize the accountability aspect (i.e how programme implementation proceeds) rather than the learning aspect of an evaluation. Formative evaluation processes, where the evaluator is given time and space to follow processes as they unfold, to establish and collect data, and to make recommendations for programme redesign, would be preferable. To promote learning, a process evaluation can assess the internal dynamics and management practices of organizations, the content and formation of their policies, their service delivery mechanisms, and so on.

The main difference between programme and process evaluations is that issues of efficiency and effectiveness can rarely be assessed conclusively before the programme has been under way for some time. Therefore, process evaluations tend to focus on whether activities are implemented according to plan and outputs are achieved on time. They also typically have fewer resources at their disposal. Comparisons are rarely undertaken, since the purpose of process evaluation is to promote internal understanding of the programme rather than to generate evidence of performance. Silver and gold methods are often too resource-demanding for use in process evaluations.

4. Gold: Using randomization as a method to gather evidence

When aiming for gold, the best possible evidence for a single study, a central question is how well one can *isolate* the impact of the programme from other factors that might affect progress towards the objective to arrive at a measurable net effect when comparing treatment and control groups. Evaluation methods using randomized data collection or experimentally based data will contribute to stronger internal validity. The use of randomization and field experiments may well be combined with bronze and silver methods so that a larger set of data, with various degrees of reliability, can inform the analysis.

This section focuses on randomization and field experiments as evaluation methods, presenting illustrative cases where data collection and analysis has benefited from an experimental approach. The primary data collection tools, surveys and interviews, are the same as for the silver and bronze methods, but they are applied differently.

4.1 The purpose of randomization

Ideally, comparison groups should start out with identical characteristics. The more similar the groups of those exposed and not exposed to the anti-corruption intervention, the better able we will be to estimate what would have happened without the programme (the *X* in box 1). A strong counterfactual scenario will provide reliable results and allow the evaluator to attribute results to the performance of the anti-corruption programme.

The most reliable method of selecting similar groups for comparison is *randomization*. By drawing randomly from a larger set of units or individuals that can be exposed or not exposed to the anti-corruption intervention, we can compose groups where members *do not differ systematically at the outset* of the programme. Any difference that subsequently arises between them can be attributed to the intervention, rather than to other factors. If used correctly in anti-corruption reform, this method will make it possible to demonstrate that observed improvements were caused by the reform programme and not by some other circumstance such as a general change in attitudes towards corruption.[17]

The example in box 2 shows how randomization can be used to assess the anti-corruption effects of regulatory reforms, in this case the monitoring of highway toll stations to reduce corruption.

[17] According to Jensen and Rahman (2011, 25), randomization also increases our ability to overcome common nonresponse and false response bias in micro-level surveys and thereby better directly measure corruption through surveys.

BOX 2: USING RANDOMIZED OBSERVATION TO MEASURE IMPACT

Consider a case in which we want to know the anti-corruption impact of placing cameras at manned toll stations. Toll booth attendants, who are civil servants, have harassed drivers by carrying out inspections on the vehicles of drivers who refuse to make an additional, informal payment to the attendant on top of the usual road fee. This has caused traffic delays and public resentment, so the government decides to invest in a monitoring regime to solve this problem. But how can we know whether the intervention has an effect?

Toll stations will be randomly assigned to two different groups: the control group and the intervention group. At the control group of toll stations there will be no cameras, and conditions will be as usual. In the intervention group, cameras will be placed near the stations. The presence of cameras is expected to influence the number of inspections conducted by toll station attendants. If cameras are recording *which* attendants are in dialogue with *which* drivers at *which* times, the attendants may perceive higher risk associated with demanding bribes.

However, in order to estimate the anti-corruption impact of placing cameras at toll stations, we must be able to estimate the levels of corruption at the stations in each group. This could be done by surveying or observing the drivers or having test drivers pass the toll stations. However it is done, it is important that the approach to measuring the level of corruption be exactly the same in the control group and the treatment group. Additionally, it will be useful to estimate the levels of corruption at randomly selected toll stations before the experiment begins in order to have a baseline to which changes in both groups over time can be compared.

If the randomization is done correctly, there should be, at the outset of the study, no systematic differences between the toll stations in the two groups. If this is the case, the differences in the corruption levels at stations where cameras have been introduced can reasonably be attributed to the intervention. This way we can measure a fairly exact anti-corruption impact of introducing cameras at toll stations in a given context. If this were undertaken as a pilot programme, it would enable a cost-benefit assessment to determine whether the costs of the reforms (e.g., purchase of cameras and monitoring of video) would outweigh the monetary benefits (fewer bribes, tantamount to a reduced tax on businesses and individuals).

If there are enough members in the treatment group, they can be randomly assigned to subgroups. If the anti-corruption initiative is designed slightly differently for each subgroup, it will be possible to test the impact of several variations of the same reform in one impact evaluation project. Using this type of experimental approach in the toll road example, design variations could include differences in the information about how the recordings will be used, differences in the stated consequences of corruption if caught, or different numbers of cameras. By varying the design of initiatives for different subgroups, we can obtain useful information about why an initiative works.[18]

Box 3 presents a fictitious case in which randomization is used to assess the impact of anti-corruption capacity-building efforts on corruption in the water sector, as well as on larger development outcomes such as water quality and coverage. In the example, one would be able to establish a causal relationship between the anti-corruption measures and changes in levels of fraud, water quality, and so forth, provided that the measures applied were the same in all local government units.

[18] The use of subgroups adds to the number of participants required to obtain statistically significant results. It is possible to randomize members at the level of groups or institutions (e.g., toll stations); it does not have to be individuals.

BOX 3: COST-EFFECTIVE USE OF RANDOMIZATION TO MEASURE IMPACT WITHIN A PROGRAMME

A local government launches a water sector reform programme with an anti-corruption component. The anti-corruption component only has funds to work with a limited number of the local government units targeted by the reform programme. Rather than "cherry picking" which units will receive anti-corruption capacity building, officials randomly assign units to this group. If baseline data are collected across the programme (and ideally also outside the programme), then the effects of anti-corruption capacity-building measures in local government units can be assessed in relation to outcomes such as the quality and coverage of water supply, number of bribes, and estimated "leakage" in water budgets due to fraud. If results are better in the areas where "corruption-proofed" local government units operate, we can attribute these changes to the impact of the anti-corruption measures. This can be done with confidence because the local government units have been randomly assigned.

However, in order to better understand *why* these measures worked, the study would benefit from qualitative research and establishing a theory of change. The example shows how evaluations can move beyond a focus on immediate corruption objectives, such as fewer informal payments or less fraud, to also document anti-corruption effects on broader development outcomes, such as better water supply.

4.2 Randomization applied

Peisakhin (2011), in a review of field experimentation and corruption, finds that very few controlled experiments have been conducted to understand the mechanisms of corruption, and hardly any have investigated what works to combat corruption. An example of carefully designed quantitative studies of the mechanisms of corruption is work by Fried, Lagunes, and Venkataramani (2010), showing that poor and socially disadvantaged people pay a higher burden than other social groups in meetings with corrupt police officers. Peisakhin concludes that field experiments on corruption help expand our knowledge and identify corruption as an obstacle to development, yet without a focus on impact they tell us little about the efficiency of policy tools.

According to Peisakhin, the majority of quantitative impact evaluations of anti-corruption describe the effect of greater information disclosure: "The link between transparency and corruption has proven to be a very promising avenue for field experimentation, because it is relatively easy to manipulate the degree to which information is disseminated" (2011, 341). Transparency-related studies fall in two categories: (a) those that relate to electoral corruption (see Wantchekon 2003; Chong et al. 2010), and (b) those that assess initiatives to combat corruption in public service delivery. The second category includes Olken's (2007) study of corruption in Indonesian villages, where the role of grassroots monitoring is evaluated. Building on these insights, it is apparent that in practice, randomized experiments have tended to focus on specific types of corruption such as electoral corruption or petty/bureaucratic corruption. The method has not been used to study political or grand corruption. Peisakhin (2011, 336) further argues that field experiments provide more relevant results than pure laboratory experiments for corruption-related issues.

Randomization has traditionally been associated with clinical trials. However, a range of other ways to randomize have been used to minimize disruption to the programmes under study. These methods are:

- *Oversubscription/lottery*. When there is limited implementation capacity for a programme or demand exceeds supply, a fair way to select beneficiaries is by lottery. This is a natural way to randomize.

- *Phase-in/piloting*. Pilots are a generally recommended way to test policies before full roll-out. Piloting can be done using random assignment, thereby creating natural comparison groups by sequencing the reforms over time.
- *Within-group randomization*. When working with, for example, facilities or government agencies, the choice does not always have to be between receiving treatment/training or not receiving it. Within the same agency, subunits can be randomly assigned to be trained using two different methods.
- *Encouragement design*. This approach can be used for policy-sensitive reform areas where one does not wish to refuse "treatment" altogether to the comparison group. Instead, one can provide special encouragement to some people, randomly chosen, to create a treatment group. This method is presented in box 4 below (Duflo, Glennerster, and Kremer 2008, 3915–18).

An area of importance to anti-corruption practitioners, where randomization can be applied easily, is how much their training activities matter in the fight against corruption. Randomization can be used to obtain a more reliable estimate of the impact of such activities because it makes it possible to establish comparison groups for evaluating training outcomes. It is, however, still critical to define the outcomes one wishes to measure. For example, it is relatively easy to assess the effectiveness of different training methods by giving participants a test before and after the training, which will show whether participants using one training method or curriculum progressed more than those using another. However, if one wishes to assess whether training leads to behavioural change, for example, or whether civil servants are less prone to take bribes after completion of a training programme, then a pre- and post-intervention assessment of bribe levels needs to be undertaken in institutions where some individuals or units have received training and others have not.

In using such an approach, it would also be useful to include a group of randomly chosen units that have been given, in addition to training, professional or monetary incentives to reduce bribe taking. In cases where one wishes to measure units rather than individuals, a cluster approach can be used, randomizing a group of subjects rather than individuals. While researchers would be concerned about having enough individuals or units to achieve statistically significant results, this would arguably not matter so much for programme officers and evaluators. The use of randomization to create similar comparison groups will always increase the validity of the results, regardless of whether one achieves significance in statistical terms.

4.3 Main hindrances

Randomization and field experiments have important advantages for the study of anti-corruption work. These methods control well for bias, which is often a major problem in evaluating the effectiveness of anti-corruption reforms. When one of these methods is used as part of an evaluation design that also uses qualitative methods such as interviews and focus groups to ground and triangulate the findings, it is hard to imagine a stronger framework. However, although randomization is considered the gold standard for evidence in individual methods, there are several reasons why it is often not the preferred choice for evaluating anti-corruption initiatives.[19]

First, randomized experiments have so far only been able to assess the impact on specific types of corruption, such as electoral corruption and petty/bureaucratic corruption, where specific practices can be tracked across a number of offices, allowing for comparison. Some anti-corruption interventions

[19] Other sectors find similar challenges in using randomization. Cook (2006) provided an early overview of this topic in the education sector.

target political and grand corruption, which has more systemic and undifferentiated impact but does not lend itself to randomized data collection.

Second, it can be difficult to generate support for these methods. Decision makers may wish to initiate reform in all units right away and may find it unacceptable not to "reform" the control group. Within donor organizations, it is important to understand that the more accurate information we have on what works, the more our development efforts will achieve in the long run. Reliable information about what works will require willingness to conduct more evaluations, including randomized and field experiments; this in turn implies that all potential beneficiaries will not benefit from reform at the same time.

Third, there are practical hindrances that may reduce the quality of results unless they are sufficiently managed (box 4). It is often difficult to establish a control group if, for example, several development agencies and the government have interventions with objectives that aim to distribute benefits across the population. If several anti-corruption programmes are under way in the same country or region, it can be difficult to isolate and assess the impact of an individual programme.

BOX 4: PRACTICAL HINDRANCES TO THE USE OF RANDOMIZATION

Consider an anti-corruption initiative that is implemented to increase reporting by staff members on corruption in their organization. How can we create separate control and treatment groups if the importance of reporting (i.e., whistleblowing) is a matter discussed by all staff and managers in the organization, as well as by the media? Will it make sense to rely on randomization for credible results of the impact evaluation effort if the two groups cannot be sufficiently separated?

The example illustrates a common difficulty that is relevant to many different anti-corruption initiatives. A way of exploiting the benefits of randomization in this context is to increase the difference between the control and treatment groups. In this example, the members of the treatment group could get some additional encouragement over and above all other anti-corruption signals. This will create a systematic difference between the treatment group, which receives specific encouragement or inducement, and the control group, which is affected only by the general awareness campaign.

The expected outcome is that there will be more reporting of corruption in both groups as a result of the anti-corruption awareness campaign, but that reporting will be higher in the treatment group, reflecting the impact of the special encouragement provided. The result will not be a perfect counterfactual estimate of an effort to increase general awareness, but it will tell us something about the value of different incentives for staff to report corruption.

A final consideration is the method's ability to provide the most *relevant* information for policymakers, as opposed to the best *evidence*. A controlled experiment requires that we be able to single out one or several priority questions for impact evaluation. However, as Ravallion (2009b, 2–3) notes, policymakers typically have a variety of questions: "Does the intervention work the way it was intended? What types of people gain, and what types lose? What proportion of the participants benefit? What happens when the programme is scaled up? How might it be designed differently to enhance impact?" Ravallion warns against putting one method (randomization) ahead of the relevant questions. Questions should dictate method, not the other way around.

So, even if the gold standard has the potential to provide more exact results than other approaches, it does not provide us with all we want to know about anti-corruption. It would be a mistake to discard all the experiences practitioners have on what works in anti-corruption on the grounds that this knowledge is not evidence-based. Field experiments using randomization can be combined with

qualitative approaches to get a better sense of the wider scope of questions that are relevant to policymakers, as explored in the section below on mixed methods.

5. Silver: Using statistical methods and advanced surveys for rigorous evidence

This section presents a large, diverse group of methods which have been labelled silver to signal that while they do not use what is considered to be the gold standard for evidence, they can provide useful and credible evidence of impact, so donors may productively use them in evaluations of anti-corruption reforms. The survey methods presented below fit circumstances and programmes where randomization for some reason is not an option. Elements of randomization can sometimes be applied, but we will now address the cases when this is typically not possible due to programme attributes.

Although silver methods can be used to evaluate a broad range of policies and programmes, they are most applicable for so-called large *n* interventions, meaning activities that target a large number of people or units (measuring, for example, individual experiences of bribery).

The following sections focus on survey-based data collection methods. The methods in section 5.1 use large data sets—obtained either through incorporation of corruption questions in household questionnaires or through bespoke micro-surveys—to conduct statistical analysis of the anti-corruption effects of an intervention.[20] These methods are often used in "quasi-experimental" designs, using data analysis rather than randomization to establish a comparison group. Section 5.2 presents methods using nonexperimental designs and standardized, advanced survey methods that often rely on finding discrepancies between data to estimate corruption, or simply construct a comparable framework for reporting individual corruption experiences at the institutional level.[21]

5.1 Statistical methods to account for anti-corruption impact

Quasi-experimental evaluation designs are rarely used in evaluations of anti-corruption activities. In some cases this is a natural consequence of the choice of evaluation question and/or lack of available data. However, if survey data can be collected for both the project and a comparison group over time, and the sample sizes are reasonably representative, then many kinds of statistical methods can be used to assess effects at the individual level. Practical issues, such as lack of baseline and longitudinal data, including for a comparison group, are often the overriding reasons why statistical methods are not used to evaluate anti-corruption efforts, rather than methodological issues (Bamberger and White 2007, 64).

[20] We will not address the many nuances that distinguish the various analytical methods. Some of the most common methods are (a) statistical matching, for example propensity score matching, (b) instrumental variables, (c) regression discontinuity designs, (d) control function methods, and (e) the difference-in-differences approach. These methods use different ways to reduce bias when establishing a comparison group through assignment in data analysis. Randomization and experiments eliminate the need for such methods, as people within the treatment and nontreatment groups are expected to have similar characteristics automatically. A useful table comparing these different methods is presented by Garcia (2011, 38–39). See also Blundell and Costa Dias (2009).

[21] If the evaluation design uses random assignment to groups or individuals, it is labelled *experimental*. If the design uses other means to create a comparison group, or multiple waves of measurement, it is called a *quasi-experimental* design. *Nonexperimental designs* do not use comparison groups or multiple waves of measurement, but they can at times still establish a counterfactual scenario through nonstatistical means. This classification is mainly concerned with the strength of the design in terms of internal validity.

The basic idea behind using statistical analysis to construct a comparison group similar to the group that has been targeted by the intervention is just as applicable to anti-corruption work as to other initiatives. The advantage of using statistical methods rather than randomization is that it is often more feasible to construct a comparison group in this way.

One of the benefits of using advanced statistical methods is the opportunity to overcome one of the principal hurdles facing anti-corruption efforts: showing the linkages between fighting corruption and achieving better development outcomes in general. Studies that identify causality, showing the impact of anti-corruption initiatives on broader development outcomes, would be valuable supplements to our anti-corruption evidence base. Box 5 shows an evaluation design using statistical matching which would allow a researcher to identify lower barriers to health care for poor people (less demand for informal payments) as a causal factor contributing to the improved health of individuals.

BOX 5: HYPOTHETICAL DESIGN USING STATISTICAL MATCHING: THE EFFECT OF ANTI-CORRUPTION WORK ON HEALTH OUTCOMES

A donor agency is implementing a health sector programme in a large region of a developing country. The programme has an anti-corruption component to reduce the practice of "informal payments"—essentially bribes—which impedes poor people from accessing services. The programme uses community monitoring as its strategy to reduce informal payments. The community monitoring is only done in health centres where civil society organizations have presence nearby to conduct the work.

Before the programme begins, household surveys are conducted across the country, establishing a baseline. These surveys ask questions which will later enable the evaluator to match all the individuals on different socioeconomic characteristics and health indicators. They also survey the respondents' use of health services. Questions on whether respondents pay informal fees, and whether these payments are considered a hindrance to the use of health services, are also included. Because a large number of people are surveyed, people with the same socioeconomic background and health indicators can be matched individually, and comparison groups can be created. Some people will live in the areas where community monitors work: they constitute the intervention group. Other people live in areas where the community monitors do not work: they are the comparison group.

Because questions are asked about informal payments, the study can isolate the effects of the community monitors using difference-in-differences analysis. If the community monitors are successful in reducing informal payments, then surveys will capture this. Furthermore, if reduced informal payments lead to (a) increased usage of health services, and (b) better health indicators, then this will be registered for each individual. Having a comparison group, with similar individuals, will provide us with a counterfactual to make sure that the increased access to health and better health indicators are due to the community monitoring and are not just a general trend.

Usually evaluations identify respondents who are using the services of a government agency undergoing reform. The intention is typically to track their level of satisfaction or trust in government, or their experience with corruption, over time. Such a pre-test/post-test design can inform us about actual changes in, for example, fraud cases, but it cannot isolate the effects of a programme. However, if the programme surveys not only the people using the services but also a comparison group, setting up a counterfactual, this allows us to see whether changes in satisfaction/trust or experiences with corruption are specific to users of the reformed services. If measured over a longer time period, such an evaluation design would be able to account for the anti-corruption impacts of the reforms, provided a satisfactory level of statistical significance can be achieved. This means that for evaluation purposes there would be benefits in sequencing the implementation of reforms, so that, for example, some

business license offices would be reformed first (treatment group) and other offices would wait (comparison group).

The major constraint for most statistical methods is the fact that the collection of survey data is time-consuming and resource-intensive. For most anti-corruption projects with small budgets, the price would be prohibitive. However, as awareness increases of the importance of anti-corruption for performance in sectors such as health, education, and water, there will be greater willingness to incorporate corruption-relevant questions into household survey questionnaires conducted for sector programmes and to conduct evaluations such as the one in box 5.[22]

Another obstacle is the fact that corruption is a sensitive topic, and questionnaires must be well designed to create valid results. The development of survey designs has improved over the past decade to better address the small-sample problems and move from perception-based measures towards more hard indicators of corrupt activity.[23] Design of questions is therefore crucial, and approaches to compensate for weak responses include indirect questioning techniques and techniques to counter respondent reticence. Methods such as direct observation, physical audits, and spot checks can also be used to minimize self-reporting bias.[24]

5.2 Standardized survey instruments

For practitioners, it is often more feasible to use standardized survey tools to document anti-corruption outcomes than to use the household surveys or bespoke micro-surveys discussed in section 5.1. Although expert advice is still needed to tailor standardized survey instruments, there is less need to reinvent the wheel. A major difference between such standardized survey instruments and other kinds of surveys is that the latter most often target people's perceptions or experiences, whereas the former also focus on outcomes not related to individuals, such as leakage of public funds or performance of facilities.

Surveys such as the Public Expenditure Tracking Survey (PETS) and Quantitative Service Delivery Survey (QSDS) are known for their diagnostic and analytical uses in measuring corruption and assessing public service delivery performance, relying on the discrepancy approach. It is less well known that these standardized surveys can be designed to assess the impact of a specific programme or reform (Gauthier 2006, iv).

The Public Expenditure Tracking Survey (box 6) is a good example of a quantitative method that does not rely on having a comparison/control group to show the effects of anti-corruption work. Essentially, the survey tracks how public money flows from central ministries to frontline agencies (schools and health facilities are the frontline services most often measured, but any agency could be used) in order to identify resource use and leakage. The sample survey methodology allows the

[22] See Campos and Pradhan (2007) and Søreide and Williams (forthcoming) for examples and debate about how corruption distorts sector performance.

[23] See Kraay and Murrell (2013) for a study of how respondents lie when they respond to corruption-related surveys.

[24] See Reinikka and Svensson (2006), Clausen, Kraay, and Murrell (2011), Banerjee, Hanna, and Mullainathan (2009), and Olken and Pande (2011) for discussions of the development of survey designs and ways to reduce bias.

evaluator to estimate how much of the originally allocated resources (financial, human, and in-kind) reach the frontline agencies, and the time it takes to deliver them (Reinikka and Svensson 2006, 3).[25]

BOX 6: EXAMPLE OF HOW TO USE PETS TO PROVE WHAT WORKS

Reinikka and Svensson used a repeat expenditure tracking survey to study the effects of improved access to public information as a tool to reduce leakage and corruption in Uganda. They used a government reform initiative on access to information as a natural experiment. The raw data suggested significant effects.

Before the reforms, only 20 per cent of school funding allocated from the central level actually reached schools. After reforms, schools received an average of 80 per cent of their annual entitlements. Statistical controls for a range of factors such as household income, teacher education, school size, and degree of supervision suggest that the information campaign can explain two-thirds of this reduction (Reinikka and Svensson 2005). As the authors themselves note, however, the study "cannot distinguish the effect of the information campaign from other policy actions or changes that simultaneously influenced all schools' ability to claim their entitlement" (Reinikka and Svensson 2003, 6). In other words, direct attribution is hard to establish, as no comparisons are made and other factors not accounted for (Hubbard 2007). Nevertheless, the surveys credibly accounts for the positive results.

Quantitative Service Delivery Surveys examine the efficiency of frontline service delivery and the waste/leakage of resources by collecting information on service providers and on relevant agents in the system (Gauthier 2006, iv). The main unit of observation is the facility in question (as opposed to, for example, a household questionnaire, which has an individual person or household as the main unit of analysis). A QSDS thus measures corruption at the input and output stages in transactions and resource flows (Duncan 2006, 153). Published data are verified against practice by means of, for example, surprise audits of facilities. This can identify both monetary fraud and nonmonetary forms of corruption such as absenteeism and ghost workers (see, for example, Reinikka and Svensson 2005).

This discrepancy method, what Olken and Pande (2011) label "graft estimation by subtraction," does not necessarily have to be applied using standardized survey instruments such as PETS and QSDS. Olken (2007) compares the recorded funds spent on a road to an independent engineering estimate of what the road should actually cost to build. This involved actually building small "test roads" to get as precise a measure as possible. Most variants of these facility surveys would be useful for measuring absenteeism or fraud but would face difficulties in measuring other variables of interests, such as how anti-corruption policies affect the quality of work.[26]

---

[25] This section draws on insights from the American Evaluation Association session, "Alternatives to the Conventional Counterfactual," held in 2009 and facilitated by Michael Bamberger, Fred Carden, and Jim Rugh. The session noted that considerations in using PETS are (a) agencies should provide a clearly defined service to the public through a large number of frontline units such as schools, clinics, public transport, and (b) the PETS analysis works best when all services are provided through a uniform set of units such as schools and when there is a uniform funding mechanism. The latter condition in some cases makes it difficult to use the method for health, as there can be a wide range of different service providers, each using different financial mechanisms. A summary of the session is available at http://www.realworldevaluation.org/uploads/Alternative_approaches_to_the_counterfactual_AEA_09.doc.

[26] Lindkvist (2012) discusses this in relation to informal payments and health workers.

Another variant of the survey method is citizen report cards (CRCs), generally used to assess public satisfaction with service delivery. They can be applied where respondents have access to the same public service facilities. There is no standard methodology for CRCs, but the approach usually entails five steps: (a) identification of issues through focus group discussions, (b) designing the survey instruments, (c) determining the sample size for the survey, (d) administration of the survey by an independent agency, and (e) collection of qualitative data (Thampi and Sekhar 2006, 236). The methodology comes from Bangalore, India, where it was developed by a group of civil society activists as a means of holding public services and utilities accountable (box 7).

People report on their experiences with public service agencies such as water, electricity, and police; how they were treated when they had a problem to resolve; whether they had to pay a bribe; and whether their problem was resolved. In their current form, report cards provide useful information on public satisfaction in general, but they are limited by the fact that comparisons are often not possible, since questions and methodology differ between approaches.

CRCs have been used mainly as a diagnostic tool to measure changes in corruption in different sectors, providing a basis for advocacy. CRCs have not been used to evaluate the impact of a specific anti-corruption reform. However, if similar units (agencies, ministries, facilities) can be identified, some units could function as the comparison group, allowing for a quasi-experimental impact evaluation design. An experienced, independent research/survey organization would ideally need to carry out the citizen report card survey if it is to be used for evaluation purposes.

The above survey tools are not easily applied, despite their somewhat standardized nature. They require substantial financial and time investments in data collection and analysis.

BOX 7: CITIZEN REPORT CARDS: THE BANGALORE EXPERIENCE

The report card on public services was created by a group of civil society institutions in Bangalore, India, in 1993. The initiative came in response to the perceived poor standards of service delivery by public institutions in the city. The civil society group had no power or influence over the authorities or the institutions delivering services in the city, so they decided that the best way to stimulate an informed debate would be to enable users of public services to give feedback on their experiences. They devised a report card that asked respondents to rate the service delivery institutions with which they had had direct contact.

The exercise produced a users' evaluation of the main service providers in the city, with each institution ranked according to its customers' reported level of satisfaction. The survey also asked detailed questions about separate aspects of service delivery (e.g., staff behaviour and quality of service provided), use of facilitation payments, and responsiveness to complaints. The results were shared with the heads of all the agencies surveyed and were given extensive coverage in the press.

The report cards had a documented impact on public awareness of the need for improvement in services delivery and were instrumental in mobilizing public pressure for improvement. This in turn triggered reform in several of the agencies that had received unfavourable ratings. A repeat survey showed significant improvement in user satisfaction with the majority of the public service institutions, proving that it is possible for independent research and pressure groups to have a positive impact on service delivery (Sundet 2004).

6. Bronze: Using evaluation methods when data, time, or budgets are lacking

Most evaluation textbooks focus on the gold and silver methods. Only a few systematically address the challenges of evaluations in situations where there is "not enough time, not enough money, or not enough data" (Bamberger, Rugh, and Mabry 2006, xxix).[27]

All anti-corruption interventions can be evaluated, but some evaluation methods are better suited to certain types of interventions and certain types of evaluation questions. The gold and silver methods presented above are all concerned with the question of impact, but the other four OECD/DAC evaluation criteria—relevance, efficiency, effectiveness, and sustainability—are equally important to policymakers.

In general, the usefulness of anti-corruption evaluations as tools for learning depends on (a) use of good evaluation methods and designs, (b) how well effects can be measured, and (c) availability of resources. These elements are interlinked. Applying the right methods will positively influence how well effects can be measured, but effect measurement also depends on operationalization of good objectives and indicators for the individual project, a process which goes beyond the scope of this paper.[28]

Arguably, a majority of anti-corruption projects still lack many of the features that are seen as beneficial for rigorous monitoring and evaluation. These include a clear logical model showing how the project is expected to contribute to reduced corruption, a baseline that includes both areas covered and not covered by the intervention (the latter to provide a counterfactual), and a system in place to produce systematic data on key indicators as part of the implementation of the project. In the absence of a sufficiently robust monitoring system and adequate resources to compensate for such gaps, we have to consider second-best solutions—and how an evaluation designed after the completion of a project can be as good as possible. Rather than attempting to force low-budget process or programme evaluations to evaluate impact, which is bound to fail, donors should develop an effective evaluation strategy in these situations, so that a combination of robust impact, outcome, and process evaluations are conducted as well as possible. This will result in a better overall evidence base and thus positively affect future programming in anti-corruption. The bronze methods presented in this section can in many cases raise the quality of evidence of such evaluations.

Evaluations typically rely on interviews, case studies, and small homemade surveys for data collection.[29] The strength of the evidence produced with these methods depends to a large extent on

---

[27] Written by Bamberger, Rugh, and Mabry (2006), *RealWorld Evaluation* is one of these exceptions and has inspired large parts of this section.

[28] The paper also does not address the policy decisions around resource allocation for evaluation purposes, beyond stating that one cannot expect an impact evaluation to be done for the price of a midterm review.

[29] *Interviews* are a popular and widely used qualitative method for anti-corruption studies. To improve the strength of the evidence, it is often valuable to combine interviews with other methods and data sources in a systematic fashion to yield a combined research methodology. This is explored in section 7 on mixed methods. C*ase study* methodological guidelines can provide possible cause-effect explanations and contextualize findings. *Surveys* are ways to collect information about a large number of people. Standardized surveys often used in the anti-corruption sector such PETS, QSDS, and CRCs are discussed in section 5.2 on standardized survey instruments. These are often costly and methodologically challenging. Evaluations could benefit from using simpler user surveys, such as satisfaction surveys or exit polls, to measure change.

the preparatory work done and on whether a systematic approach to data collection and analysis has been taken.

Time and money may be additional constraints. Many donors and government agencies do not focus on monitoring and evaluation until late in the project cycle. The most common reaction once the demands for evidence appear is to commission a midterm review or end-of-programme review, conducted over a short period. Although evaluators carrying out such hasty reviews are often asked to assess impact, the reviews often largely ignore issues of rigour and validity of conclusions, as they are not designed to provide substantiated evidence on outcomes, let alone impacts, of an intervention.

This section addresses common concerns about how to produce credible evidence in situations where there is no baseline data, the logical model is unclear, indicators and data are lacking, time constraints are present, and the evaluation budget is relatively low. We discuss how anti-corruption practitioners faced with some or all of these constraints can use the following methods:

- Rapid or alternative data collection
- Reconstruction of baseline or comparison data
- New technologies, social media, and other secondary data

6.1 Rapid or alternative data collection methods

To solve the problem of missing data, or when proper data collection would be too costly, anti-corruption practitioners can use a number of alternatives to the traditional, labour-intensive survey questionnaires. Cost-effective methods for data collection include more targeted instruments such as structured observation, exit surveys, focus group discussions, and rapid surveys. To make the most of scarce data, evaluation practitioners will often triangulate different types of data (interviews, secondary documentation, statistics, etc.). The mixed-methods approach and triangulation method are further explained in section 7.

Structured observation, covering situations where corrupt transactions are known to take place, can be a cost-effective way of collecting evidence. Examples include innovative use of photography or video or witness observation, all of which can be less costly than surveys and more reliable in documenting a hidden phenomenon like corruption.

Duflo, Hanna, and Ryan (2011) provide an example of photography used in structured observation. They conducted a randomized study on teacher absenteeism, which has been described as a form of "quiet corruption" (World Bank 2010). Each teacher was given a camera, along with instructions to have one student take a picture of the teacher and the class at the start and close of each school day. The study was able to present strong evidence for the effect of monitoring. Other approaches using witness observation are provided in box 8.

Exit surveys, also known as exit polling, are a way to obtain rapid feedback on, for example, the usefulness of public hearings, or find out whether people have paid bribes at a specific institution. So-called *rapid surveys* are very short questionnaires that ask only a few questions, reducing the time needed to collect and analyse data. *Web-based and electronic surveys* can also be used as an alternative to the traditional resource-intensive surveys in areas where there is access to the technology (see section 6.3).

Focus groups can be established easily and quickly if one uses market research companies to construct the groups. A variant of focus groups, community interviews, can be conducted using participatory rural appraisal methods at a lower cost than surveys.

Purposeful sampling techniques can be used to reduce the number of interviews needed to assess people's experiences with corruption, evaluate use of and satisfaction with public services, and so forth. For example, rather than interviewing a random sample of users of a facility, it can make sense to draw a "critical sample" of corruption victims only, but only if one is trying to answer questions the characteristics of such victims, not the frequency of victimisation (Bamberger, Rugh, and Mabry 2006, 270–71). In other cases it might be better to conduct random spot checks, for example, to assess levels of compliance within government units with anti-corruption measures. Randomization in this case enables the evaluator to produce credible evidence on compliance levels without having to analyse all government units.

BOX 8: EXAMPLES OF WITNESS OBSERVATION

Olken and Barron (2009) provide data on bribes truck drivers paid to police on their routes to and from the Indonesian province of Aceh. Enumerators accompanied truck drivers as passengers and noted what the truck drivers paid each time they were stopped at a police checkpoint or weigh station. Over approximately 300 trips, they observed more than 6,000 illegal payments. The total cost of corruption constituted 13 per cent of the marginal cost of the trip.

Sequeira and Djankov (2010) used observation in their study of customs corruption in Mozambique and South Africa. They shadowed clearing agents who process customs for cargo as it passes through the ports, enabling them to directly observe bribe payments to port and border post officials for a random sample of 1,300 shipments. The study showed that bribes account for, on average, 14 per cent of shipping costs for a standard container in Maputo, Mozambique, and 4 per cent in Durban, South Africa.

6.2 Reconstructing baseline or comparison data

In cases where a baseline has not been established for the anti-corruption programme, there are ways to reconstruct it using forms of retrospective analysis. It is sometimes possible to find secondary data that contain enough information to create a credible baseline. One can also ask people to think back and report their perception of the baseline situation. However, such recall has many limitations, and it is worth questioning what the value of the exercise will be in each case. The quality of recall-based data can be affected by respondents' intentional suppression of the facts: respondents may withhold accurate details because they have committed criminal activity or fear reprisals from others, or they may expect to benefit from distorting the truth. Sometimes political or ethnic affiliations can encourage respondents to downplay the success of specific anti-corruption reforms. Suppression of the facts can also be unintentional, due to poor memory or nostalgia. For various reasons respondents might either underestimate or overestimate facts, such as the number of bribes they had to pay to a specific ministry in the past.

If a "before-and-after" analysis is not possible because no baseline is available, a compromise is to do "with-and-without," comparing the intervention group with a nonintervention group at the end of the programme only. Where baseline data exist but there is no comparison group (because the programme only concerned itself with direct beneficiaries, e.g., the users of services, not the nonusers), then secondary data can at times be used to construct a comparison group through statistical matching, even later in the programme lifetime.

The issue of reconstructing baseline or comparison data can be crucial in cases where donors must decide whether a supposedly successful pilot initiative should be scaled up, but no documentation has been collected. The hypothetical example of a community monitoring pilot programme is described in box 9.

Having to reconstruct baseline and/or comparison data is far from ideal, and should be avoided if better options exist. Much time and effort can be saved by collecting data proactively, which means devoting attention to indicator development and data collection at the programme design stage. However, if no other options exist, it is usually better to use retrospective analysis to generate some added "approximate evidence" rather than to have none at all.

BOX 9: INVESTIGATING SUCCESS IN COMMUNITY MONITORING

A hypothetical pilot project working with community groups in the construction sector is seen as highly successful in reducing corruption, and several stakeholders want to scale up the project across the country. Unfortunately, no baseline was established for the level of corruption (defined as both leakage and fraud) in the construction sector in the areas where the community monitors worked. Before investing in an expensive scale-up of the initiative, therefore, donors request a retrospective evaluation of the initiative.

Fortunately, a baseline can be reconstructed. The evaluation team analyses and cross-checks audit reports, government expenditure budgets, and company records and conducts spot checks to estimate whether construction projects were finished to the agreed standard. In so doing, the team finds that the level of funds that are unaccounted for has remained stable during the past 10 years, including the five years the community monitoring programme has been in existence. When they compare the project region with other regions in the country where no anti-corruption initiatives were in place, the picture is the same. In short, no evidence can be found to back up the "success" story. Donors are not happy that the pilot was not a success, but they are pleased that they decided to invest in evidence before making a decision to scale up the project.

6.3 New technologies, social media, corruption indices, and other secondary data

To save on costs relating to primary data collection, one can often make use of secondary sources of data. The quality of data is increased when primary and secondary sources of data are combined and triangulated (see White and Bamberger 2008, 8).

In areas where information and communications technology (ICT) is widely used, social media, mobile phones, and Internet surveys can be useful sources of information. The right type of information can be used to estimate trends in corruption, which can show "real-time" effects of initiatives. Potentially useful data sources include crowd-sourcing initiatives that document actual cases of corruption and identify points in the system where bribery occurs.

A caveat is that the information obtained from such ICT tools is often difficult to verify and can potentially be manipulated. For example, since no names are given on social network sites, it is impossible to verify the anonymous reports. For this reason, the data should not be used as the only source of evidence, but should complement other data sources. The point is, however, that the quality of the results in these ex-post evaluations depends on creativity. Just as businesses and governments benefit from the use of such real-time, crowd-sourcing data for user feedback, for example, donor programmes can use these technologies to understand the impact of anti-corruption initiatives.

One of the better-known examples of crowd-sourcing in the context of corruption is the website Ipaidabribe.com (box 10). This type of initiative is usually used as a tool for advocacy; however, the data can also be used for evaluation purposes. Using data from a source like Ipaidabribe.com, one could study whether a newly introduced measure has resulted in a decrease of corruption reports over

time. The financial costs of producing these data are negligible, and the time needed for data analysis is also minimal.

Moreover, such Internet-based, self-reporting surveys are not only useful for bribery cases. Complaints of any kind could be reported. Donor programme monitoring and evaluation systems could often benefit from having an easily accessible grievance reporting system to identify red flags for corruption and fraud. ICT tools, whether using the Internet or mobile text messaging, can provide managers with direct feedback from beneficiaries. Although the data would be too "noisy" to serve as the only source of evidence, the information will still be useful for evaluators and managers.

Use of data from ICT sources is, however, rare in evaluations of anti-corruption programmes. The most common sources of secondary data used for this purpose are corruption indices such as Transparency International's Corruption Perceptions Index or the World Bank's Control of Corruption Index. Both are perceptions-based indices and may not measure corruption accurately (Olken and Pande 2011, 4). Moreover, such broad countrywide indices are often inappropriate for policy or programme evaluations (Søreide 2006; Arndt and Oman 2006; Knack 2006). They work better as advocacy tools or as general measurements of corruption trends. A few indices offer the option of disaggregating data to the institution or sector level, which in some cases enables the data from these indices to be used as part of evaluations. An example is Transparency International's Global Corruption Barometer (Johnsøn and Hardoon 2012) for the public sector, and World Bank Enterprise Surveys for the private sector.

Perhaps due to the easy availability of the corruption indices, evaluations tend to use few other sources of secondary data. However, some *government administrative records* have proven to contain useful information. Sometimes the story of corruption can be gleaned from the gap between official and actual expenditures. For example, Niehaus and Sukhtankar (2010) compare government administrative data with independent survey data and find significant overreporting of days and underpayment of wages in India's employment guarantee scheme. Other sources of information could be court records, police files, or customs papers. Poor record-keeping standards and pervasive reporting bias are evident obstacles, and they have to be assessed on a case-by-case basis. Verifying the information with records from newspapers, community organizations, universities, and other sources is helpful.

To conclude, bronze methods include creative ways of performing good-quality evaluations even when the programme to be evaluated has not been designed with evaluation in mind. These methods can be used to deal with missing baselines and comparative data, providing good-quality evaluations at low cost while contributing to better understanding of what works in anti-corruption.

BOX 10: CROWD-SOURCING FOR ANTI-CORRUPTION

The website Ipaidabribe.com was set up in India in 2010 to collect and publicize anonymous reports of bribes paid, bribes requested but not paid, and requests that were expected but did not occur. An example of how to use such data to identify risk areas and drum up support for reforms is provided by the transport commissioner for the state of Karnataka. Using Ipaidabribe.com, citizens posted reports on corruption in the system for issuing driving and vehicle licenses. As a result, authorities changed the system to reduce corruption risks by introducing, among other reforms, online applications for driving licenses and video surveillance of motor vehicle inspectors.*Source*: Strom 2012.

7. Mixed methods and triangulation: Principles of strong evaluation design

Although this paper focuses on methods for anti-corruption evaluation and research and not on evaluation design, the principle of mixed methods for evaluation design is important and will be briefly explained.

Mixed methods refers to the combination of quantitative and qualitative methods in an iterative and complementary way. As argued in this paper, some methods are more likely to yield reliable, replicable, and valid findings than others. But in terms of an overall evaluation design, a well-considered combination of methods will often provide the best possible result. Combining qualitative and quantitative methods is particularly useful for analysis of complex social questions such as those relevant to anti-corruption work.

Although the benefits of mixing methods are widely recognized, it is rarely done in real-life evaluations (Bryman 2006). This gap between prescription and practice is a result of several factors. Some researchers and evaluators have a tendency to work with methods they are familiar with rather than those that are most relevant for answering the research question, and funding rarely allows for multidisciplinary teams.

Bamberger, Rao, and Woolcock (2010) provide a good general introduction to the use of mixed methods, and Johnsøn et al. (2011, 40–41) discuss the benefits of a mixed-methods design for corruption measurement and anti-corruption evaluations.

Triangulation can be understood to mean the use of more than one method or source of data in the study of a social phenomenon (Bryman 2004, 275). Although the term is often used interchangeably with mixed methods, triangulation is actually a broader principle, focusing not just on design but also on the analysis and interpretation of data.

Triangulation sheds light on issues from different angles to "overcome the problems that stem from studies relying upon a single theory, a single method, a single set of data [. . .] and from a single investigator" (Mikkelsen 2005, 96). All methods benefit from having their findings triangulated, or cross-checked, with other data sources to increase the validity of evidence. Because corruption is a complex social construct, neither one indicator nor one analytical method would normally be sufficient to explain corruption. Accordingly, triangulation of both methods and indicators is recommended. Triangulation of indicators is done to reduce the threat of low construct validity. In sum, evidence is often stronger if supported by several methods (both qualitative and quantitative), and more valid if more than one relevant indicator measures it.

It is difficult to provide recommendations on evaluation designs for anti-corruption activities without knowing the nature of the evaluation question, the programme attributes, how people are exposed to the intervention, and the available data. In most cases, however, a combination of quantitative and qualitative methods will result in a stronger evaluation.

8. Conclusion

A wide range of evaluation methods are available that can document the outcomes and impact of anti-corruption interventions. Some methods produce stronger evidence than others, but all of the methods presented in this paper can contribute in some way to building a better evidence base for what works and why. Advanced evaluation methods are rarely used in donor-funded evaluations of anti-corruption interventions. The examples provided come mainly from academic research. However, with careful planning and adequate resources, these methods can and should be used to evaluate donor-funded programmes. Learning needs to come both from academic research and from donors' own evaluations. The best results can be achieved if the evaluation is designed and planned before the programme begins. This would bring far more useful results than the often-preferred quick, ex-post assessment of programme performance done by means of midterm and end-of-term reviews.

Anti-corruption interventions come in many different shapes and forms. Some target individual behavioural change, while others aim to change government processes, laws, or policies. Not all objectives lend themselves to statistical analysis, and not all programmes merit the substantial investment of an impact evaluation. But some do, and a selection of such programmes should be rigorously evaluated. As shown in the examples above, many forms of corruption can be credibly measured, and many outcomes can be documented. The evaluation questions, programme attributes, and available data should determine the evaluation design and method, not the other way around. Only policymakers and programme staff can identify which evaluation questions matter most to them. The point is that many questions can in fact be answered if only the proper investment in evaluation is made.

References

Andrews, Matt. 2013. *The Limits of Institutional Reform in Development: Changing Rules of Realistic Solutions*. Cambridge, UK: Cambridge University Press.

Arndt, Christine, and Charles Oman. 2006. *Uses and Abuses of Governance Indicators*. Paris: OECD Development Centre.

Bai,J., S. Jayachandran, E. J. Malesky and B. A. Olken. 2013. *Does Economic Growth Reduce Corruption? Theory and Evidence from Vietnam*. Unpublished manuscript per July 2013 (available online).

Bamberger, Michael. 2009. "Why Do Many International Development Evaluations Have a Positive Bias? Should We Worry?" *Evaluation Journal of Australasia* 9, no. 2: 39–49.

Bamberger, Michael, Vijayendra Rao, and Michael Woolcock. 2010. "Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development." In *Handbook of Mixed Methods Research*, 2nd ed., edited by Abbas Tashakkori and Charles B. Teddlie, 613–42. Thousand Oaks, CA: Sage.

Bamberger, Michael, Jim Rugh, and Linda Mabry. 2006. *RealWorld Evaluation: Working under Budget, Time, Data, and Political Constraints*. Thousand Oaks, CA: Sage.

Bamberger, Michael, and Howard White. 2007. "Using Strong Evaluation Designs in Developing Countries: Experience and Challenges." *Journal of MultiDisciplinary Evaluation* 4, no. 8: 58–73.

Banerjee, Abhijit, Rema Hanna, and Sendhil Mullainathan. 2009. *Corruption.* MIT Department of Economics Working Paper 12-08; HKS Working Paper 12-023. Cambridge, MA: MIT Department of Economics.

Barder, Owen. 2012. "Complexity, Adaptation and Results." *Global Development: Views from the Center* (blog, Center for Global Development). http://blogs.cgdev.org/globaldevelopment/2012/09/complexity-and-results.php.

Basu, Kaushik. 2011. "Why, for a Class of Bribes, the Act of Giving a Bribe Should Be Treated as Legal." Memo, Indian Ministry of Finance, New Delhi.

Björkman, Martina, and Jakob Svensson. 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda." *Quarterly Journal of Economics* 124, no. 2: 735–69.

Blundell, Richard, and Monica Costa Dias. 2009 "Alternative Approaches to Evaluation in Empirical Microeconomics." *Journal of Human Resources* 44, no. 3: 565–636.

Bryman, Alan. 2004. *Social Research Methods*. 2nd ed. Oxford, UK: Oxford University Press.

———. 2006. "Integrating Quantitative and Qualitative Research: How Is It Done?" *Qualitative Research* 6, no. 1: 97–113.

Campos, J. Edgardo, and Sanjay Pradhan, eds. 2007. *The Many Faces of Corruption: Tracking Vulnerabilities at the Sector Level.* Washington, DC: World Bank.

Chong, Alberto, Ana L. De La O, Dean Karlan, and Leonard Wantchekon. 2010. "Information Dissemination and Local Governments' Electoral Returns: Evidence from a Field Experiment in Mexico." Unpublished paper, Yale University, New Haven, CT.

Clapham, Christopher. 1985. *Third World Politics: An Introduction.* Routledge.

Clausen, Bianca, Aart Kraay, and Peter Murrell. 2011. "Does Respondent Reticence Affect the Results of Corruption Surveys? Evidence from the World Bank Enterprise Survey for Nigeria." In *International Handbook on the Economics of Corruption*, vol. 2, edited by Susan Rose-Ackerman and Tina Søreide. Cheltenham, UK: Edward Elgar.

Cook, Thomas. 2006. "Describing What Is Special about the Role of Experiments in Contemporary Educational Research? Putting the 'Gold Standard' Rhetoric into Perspective." *Journal of MultiDisciplinary Evaluation* 3, no. 6: 1–7.

Cosgrave, John, Ben Ramalingam, and Tony Beck. 2009. *Real-Time Evaluations of Humanitarian Action.* London: Overseas Development Institute.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, vol. 4, edited by T. Paul Schultz and John Strauss. Amsterdam: Elsevier.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102, no. 4: 1241–78.

Duncan, Nick. 2006. "The Non-Perception Based Measurement of Corruption: A Review of Issues and Methods from a Policy Perspective." In *Measuring Corruption*, edited by Charles Sampford, Arthur Shacklock, Carmel Connors, and Fredrik Galtung, 131–62. Aldershot, UK: Ashgate.

Fried, Brian J., Paul Lagunes, and Atheendar Venkataramani. 2010. "Corruption and Inequality at the Crossroads: A Multimethod Study of Bribery and Discrimination in Latin America." *Latin American Research Review* 45, no. 19: 76–97.

Funnell, Sue, and Patricia Rogers. 2011. *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. San Francisco: Jossey-Bass.

Garcia, Melody. 2011. *Micro-Methods in Evaluating Governance Interventions*. Discussion Paper 7/2011. Bonn: German Development Institute.

Gauthier, Bernard. 2006. *PETS-QSDS in Sub-Saharan Africa: A Stocktaking Study*. Washington, DC: World Bank.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2011. *Impact Evaluation in Practice*. Washington, DC: World Bank.

Hanna, R., S. Bishop, S. Nadel, G. Scheffler, and K. Durlacher. 2011. *The Effectiveness of Anti-Corruption Policy: What Has Worked, What Hasn't, and What We Don't Know—A Systematic Review*. Technical Report. London: EPPI-Centre, Social Science Unit, Institute of Education, University of London.

Hubbard, Paul. 2007. *Putting the Power of Transparency in Context: Information's Role in Reducing Corruption in Uganda's Education Sector.* Working Paper 136. Washington, DC: Center for Global Development.

Hunt, Jennifer. 2006. "Why Are Some Public Officials More Corrupt than Others?" In *International Handbook on the Economics of Corruption*, edited by Susan Rose-Ackerman, 323–51. Cheltenham, UK: Edward Elgar.

IEG (Independent Evaluation Group). 2011. *High Impact Evaluations: Exploring the Potential of Real-Time and Prospective Evaluations*" Summary of a workshop, 27 January. Washington, DC: World Bank.

Jensen, Nathan M., and Aminur Rahman. 2011. *The Silence of Corruption: Identifying Underreporting of Business Corruption through Randomized Response Techniques*. Policy Research Working Paper 5696. Washington, DC: World Bank.

Johnsøn, Jesper. 2012. *Theories of Change in Anti-Corruption Work: A Tool for Programme Design and Evaluation*. U4 Issue 2012:6. Bergen, Norway: U4 Anti-Corruption Resource Centre.

Johnsøn, Jesper, and Deborah Hardoon. 2012. *Why, When and How to Use the Global Corruption Barometer*. U4 Brief 2012:5. Bergen, Norway: U4 Anti-Corruption Resource Centre.

Johnsøn, Jesper, Hannes Hechler, Luís De Sousa, and Harald Mathisen. 2011. *How to Monitor and Evaluate Anti-Corruption Agencies: Guidelines for Agencies, Donors, and Evaluators.* U4 Issue 2011:8. Bergen, Norway: U4 Anti-Corruption Resource Centre.

Johnsøn, Jesper, Nils Taxell, and Dominik Zaum. 2012. *Mapping Evidence Gaps in Anti-Corruption: Assessing the State of the Operationally Relevant Evidence on Donors' Actions and Approaches to Reducing Corruption*. U4 Issue 2012:7. Bergen, Norway: U4 Anti-Corruption Resource Centre.

Johnston, Michael. 2005. *Syndromes of Corruption: Wealth, Power, and Democracy*. Cambridge, UK: Cambridge University Press.

Kaufmann, Daniel. 1997. "Corruption: Some Myths and Facts." *Foreign Policy*, Summer, 114–31.

Khan, M.H. 2012. Governance during Social Transformations: Challenges for Africa. *New Political Economy*. 17(5): 667-675.

Klitgaard, Robert. 1988. *Controlling Corruption*. Berkeley: University of California Press.

Knack, Stephen. 2006. *Measuring Corruption in Eastern Europe and Central Asia: A Critique of the Cross-Country Indicators*. Policy Research Working Paper 3968. Washington, DC: World Bank.

Kraay, Aart, and Peter Murrell. 2013. *Misunderestimating Corruption*. Policy Research Working Paper 6488. Washington, DC: World Bank.

Lambsdorff, Johann. 2006. "Causes and Consequences of Corruption." In *International Handbook on the Economics of Corruption*, edited by Susan Rose-Ackerman, 3–51. Cheltenham, UK: Edward Elgar.

Lindkvist, Ida. 2012. "Informal Payments and Health Worker Effort: A Quantitative Study from Tanzania." *Health Economics*, 27 November. doi: 10.1002/hec.2881.

Liverani, Andrea, and Hans E. Lundgren. 2007. "Evaluation Systems in Development Aid Agencies." *Evaluation* 13, no. 2: 241–55.

Mikkelsen, Britha. 2005. *Methods for Development Work and Research: A New Guide for Practitioners*. Thousand Oaks, CA: Sage.

Niehaus, Paul, and Sandip Sukhtantar. 2010. *Corruption Dynamics: The Golden Goose Effect*. BREAD Working Paper 223. Bureau for Research and Economic Analysis of Development.

Norad (Norwegian Agency for Development Cooperation). 2011. *Joint Evaluation of Support to Anti-Corruption Efforts 2002–2009.* Oslo: Norad.

OECD/DAC (Organisation for Economic Co-operation and Development, Development Assistance Committee). 2010a. *Glossary of Key Terms in Evaluation and Results Based Management.* Paris: OECD.

———. 2010b. *Quality Standards for Development Evaluation.* Paris: OECD.

Olken, Benjamin. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115, no. 2: 200–49.

Olken, Benjamin A., and Patrick Barron. 2009. "The Simple Economics of Extortion: Evidence from Trucking in Aceh." *Journal of Political Economy* 117, no. 3: 417–52.

Olken, Benjamin, and Rohini Pande. 2011. *Corruption in Developing Countries.* NBER Working Paper 17398. Cambridge, MA: National Bureau of Economic Research.

Peisakhin, Leonid V. 2011. "Field Experimentation and the Study of Corruption." In *International Handbook on the Economics of Corruption*, vol. 2, edited by Susan Rose-Ackerman and Tina Søreide. Cheltenham, UK: Edward Elgar.

Pope, Jeremy. 2000. *Confronting Corruption: The Elements of a National Integrity System*. Berlin: Transparency International.

Pritchett, Lant. 2002. "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *Policy Reform* 5, no. 4: 251–69.

Pritchett, Lant, Michael Woolcock, and Matt Andrews. 2010. *Capability Traps? The Mechanisms of Persistent Implementation Failure*. Working Paper 234. Washington, DC: Center for Global Development.

Ravallion, Martin. 2009a. "Evaluating Three Stylised Interventions." *Journal of Development Effectiveness* 1, no. 3: 227–36.

———. 2009b. "Should the Randomistas Rule?" *Economists' Voice* 6, no. 2 (February).

Reinikka, Ritva, and Jakob Svensson. 2003. *Survey Techniques to Measure and Explain Corruption*. Washington, DC: World Bank.

———. 2005. "Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda." *Journal of the European Economic Association* 3, no. 2–3: 259–67.

———. 2006. "Using Micro-Surveys to Measure and Explain Corruption." *World Development* 34, no. 2: 359–70.

Rose-Ackerman, Susan. 1978. *Corruption: A Study in Political Economy.* New York: Academic Press.

———. 1999. *Corruption and Government: Causes, Consequences, and Reform*. Cambridge, UK: Cambridge University Press.

Rose-Ackerman, S. and R. Truex. 2013. "Corruption and Policy Reform." In B. Lomborg (ed.) *Global Problems, Smart Solutions: Costs and Benefits*. Cambridge, UK: Cambridge University Press.

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach*. 7th ed. Thousand Oaks, CA: Sage.

Rothstein, Bo. 2011. "Anti-Corruption: The Indirect 'Big Bang' Approach." *Review of International Political Economy* 18, no. 2: 228–50.

Savedoff, W.D., R. Levine, and N. Birdsall. 2006. *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Washington, DC: Center for Global Development.

Sequeira, Sandra, and Simeon Djankov. 2010. *An Empirical Study of Corruption in Ports*. MPRA Paper 21791, Munich Personal RePEc Archive. http://mpra.ub.uni-muenchen.de/21791/.

Søreide, Tina. 2006. *Is It Wrong to Rank? A Critical Assessment of Corruption Indices*. CMI Working Paper 2006:1. Bergen, Norway: Chr. Michelsen Institute.

Søreide, Tina, and Aled Williams, eds. 2014. *Grabbing Development: Real World Challenges.* Cheltenham and Northhampton, UK: Edward Elgar Publishing. Forthcoming.

Stern, Elliot, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, and Barbara Befani. 2012. *Broadening the Range of Designs and Methods for Impact Evaluations*. Working Paper 38. London: Department for International Development.

Strom, Stephanie. 2012. "Web Sites Shine Light on Petty Bribery Worldwide." *New York Times*, March 6.

Sundet, Geir. 2004. *Public Expenditure and Service Delivery Monitoring in Tanzania: Some International Best Practices and a Discussion of Present and Planned Tanzanian Initiatives*. Working Paper 04.7. Dar es Salaam: HakiElimu.

Svensson, Jakob. 2005. "Eight Questions about Corruption." *Journal of Economic Perspectives* 19, no. 3: 19–42.

Thampi, Gopakumar, and Sita Sekhar. 2006. "Citizen Report Cards." In *Measuring Corruption*, edited by Charles Sampford, Arthur Shacklock, Carmel Connors, and Fredrik Galtung, 233–50. Aldershot, UK: Ashgate.

Treisman, Daniel. 2000. "The Causes of Corruption: A Cross-National Study." *Journal of Public Economics* 76, no. 3: 399–457.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55, no. 3: 399–422.

White, Howard. 2006. *Impact Evaluation: The Experience of the Independent Evaluation Group*. Washington, DC: World Bank.

———. 2009. *Theory-Based Impact Evaluation: Principles and Practice*. 3ie Working Paper 3. New Delhi: International Initiative for Impact Evaluation.

White, Howard, and Michael Bamberger. 2008. "Introduction: Impact Evaluation in Official Development Agencies." *IDS (Institute of Development Studies) Bulletin* 39, no. 1: 1–11.

White, Howard, and Daniel Phillips. 2012. *Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework*. Working Paper 15. New Delhi: International Initiative for Impact Evaluation.

World Bank. 2010. *Africa Development Indicators 2010: Silent and Lethal: How Quiet Corruption Undermines Africa's Development Efforts.* Washington, DC: World Bank.

U4 Anti-Corruption Resource Centre
Chr. Michelsen Institute (CMI)
Phone: +47 47 93 80 00
Fax: +47 47 93 80 01
u4@u4.no
www.U4.no

P.O.Box 6033 Bedriftssenteret
N-5892 Bergen, Norway
Visiting address:
Jekteviksbakken 31, Bergen

This U4 Issue is also available at:
www.u4.no/publications

INDEXING TERMS:
Corruption
Anti-corruption
Governance
Evaluation
Monitoring
Impact
Result chain
Mixed methods
Theory of change
Evidence
Learning

Cover image by
Jesper Johnsøn
U4/CMI

CMI CHR. MICHELSEN INSTITUTE

Evaluations of donor-funded anti-corruption reforms and programmes would benefit from upgrading and diversifying the methods used to document effects. Better evaluations in turn would improve the evidence base for the effectiveness of specific anti-corruption interventions. Using real and hypothetical examples, this paper offers practical guidance to practitioners who design, implement, and disseminate evaluations and research on anti-corruption. A range of quantitative and qualitative methods can be used to answer operational questions on the impact of anti-corruption interventions. Some methods can produce stronger evidence than others for a specific evaluation, but there are trade-offs between rigour and costs, and between aspiration and feasibility. Donors should let the evaluation question, programme attributes, and availability of data determine the most appropriate methods for a given study. With careful planning and adequate resources, donors can use many of the methods presented in this paper. This should give more reliable results and produce needed knowledge on what works in anti-corruption, laying the basis for more effective anti-corruption initiatives.

U4 Anti-Corruption Resource Centre
www.U4.no