Evaluation Strategies for Global and National Measures Against Illicit Financial Flows

Michael G. Findley

Increased attention to problems of corruption, organized crime, and terrorism has led to greater international focus on illicit financial flows (IFFs), including in the form of generating strategies for combating such flows. At least four UN conventions—the Convention Against Illicit Traffic in Narcotic Drugs and Psychotropic Substances, the International Convention for the Suppression of the Financing of Terrorism, the Convention Against Transnational Organized Crime, and the Convention Against Corruption—have been adopted, as have many regional agreements, all of which seek to stem IFFs. UN Sustainable Development Goal target 16.4, moreover, calls for addressing IFFs as a critical priority for the developing world.[1]

While global and national measures to combat IFFs are diverse and advanced by many organizations, they were most prominently developed through the Forty Recommendations of the Financial Action Task Force (FATF).[2] Governments have delegated standard-setting authority to FATF, and nearly all countries in the world have agreed, in principle, to abide by FATF standards, either because the countries are FATF members or belong to one of FATF's nine regional satellite organizations.[3] National adoption and enforcement, though, vary widely. Among others, the World Bank and the International Monetary Fund consider FATF to be the authoritative organization for setting and enforcing anti–money laundering (AML) standards.[4] The range of global standards is impressively broad; a few critical measures, including the establishment of beneficial ownership and the implementation of AML recommendations, are highlighted below.

Evaluation strategies for these global and national measures against IFFs have been severely neglected, in contrast to issues of IFF measurement. Much has been written on whether IFFs are increasing or decreasing year to year, with at least implicit attribution to existing global and national measures. Extant attributions of success should be interpreted with caution because existing research approaches are not well suited to establish confidently that current policies are responsible for any shifts and that alternative explanations are not responsible. This argument can be contextualized through a brief discussion of different evaluation methods, including of what evaluation experts count as credible evidence.

National statutory compliance with international standards does not correlate much with actual compliance; knowledge of international standards does not motivate compliance with the mandate to establish beneficial ownership; the threat of national enforcement does appear to motivate compliance; the risk-based approach in know-your-customer (KYC) process appears to motivate greater compliance with international standards; and although tax havens have a bad reputation, they appear to be among the most compliant jurisdictions in the world.[5]

Evaluating IFFs defined in the macroeconomic sense of aggregating unlawful cross-border movements of money (assuming the severe measurement issues can be remedied) is unlikely to yield valuable conclusions to inform specific policy recommendations. Instead, scholars and policymakers should prioritize more rigorous evaluations of specific programs applied to narrower illicit flows. Rather than focus so much attention on aggregate flows, evaluators should examine the efficacy of global governance strategies considering meso- or micro-level dimensions of the IFFs.

MEASUREMENT AND EVALUATION OF ILLICIT FINANCIAL FLOWS

IFFs take on many forms. Much of the literature emphasizes fraudulent misinvoicing of trade to the developing world.[6] Substantial effort has been devoted to understanding how money associated with such trade, or more broadly for transferring bulk cash or otherwise, is concealed or laundered. Untraceable shell corporations with bank accounts (so that the accounts are de facto anonymous) provide likely the most common mechanism for doing so, whether in the specific domains of money laundering, transnational corruption, tax evasion, or other related crimes.[7] The FATF recommendations are designed, in particular, to prevent such money laundering.

The scrutiny paid to fraudulent misinvoicing of trade is likely due to the attention that Global Financial Integrity (GFI) and other advocacy nongovernmental organizations have brought to this form of IFFs. While trade misinvoicing is potentially extremely consequential, with global estimates consistently on the order of trillions of dollars annually, estimating the scale of these activities is a precarious enterprise and existing estimates could be highly inaccurate.[8] Maya Forstater has produced several excellent treatments of the measurement problems, including in this collection.[9] Michael Levi, Peter Reuter, and Terence Halliday contend that no one takes the GFI estimates seriously in the sense of using the estimates for evaluating policy effectiveness and instead argue that the estimates are mere advocacy claims.[10] Without solving, or at least mitigating, the nontrivial measurement challenges, evaluation at this macroeconomic level is near impossible.[11]

Outside the measurement problems currently under debate, evaluation of global governance strategies to stem IFFs has been almost nonexistent. Levi and coauthors compellingly argue that evaluation in the AML and IFF space is miserable at best. They provide a scathing critique of the state of evaluation, perhaps best captured in their claim that "despite the publication of national Mutual Evaluation Reports (MERs) and, more recently, National Risk Assessments, the fact is that there has been minimal effort at AML evaluation, at least in the sense in which evaluation is generally understood by public policy and social science researchers, namely how well an intervention does in achieving its goals."[12] They further note: "The ideal evaluation would take some measure of the target activity, such as the total amount of money laundered, and estimate how much that has been reduced by the imposition of AML controls." They demonstrate just how little data has actually been used to evaluate AML efforts. A broader review of the scholarly literature provides no greater cause for optimism. Outside a handful of isolated studies, little rigorous research provides any basis for evaluation.[13] And yet, as J. C. Sharman has noted, the global AML regime has developed and spread quickly even without any evidence of suitability or success.[14]

EVALUATION STRATEGIES

The literature on evaluation methods primarily distinguishes between performance and impact evaluations; performance focuses on how a program has been implemented and impact evaluation focuses on the effects of the program. Of course, the two are not always easily separable. The FATF Mutual Evaluation Reports examine both whether global standards are being implemented nationally (performance component) and whether those national laws are effective in practice (impact component).

Impact evaluations measure the change in some outcome (e.g., IFFs) that is attributable to a specific, defined intervention (e.g., KYC rules). Because the task is to make proper attribution, one must be able to demonstrate that the intervention is responsible for the change in outcome, which requires ruling out alternative explanations that could confound the inferences made. The best practice for ruling out alternative explanations is to construct rigorously a counterfactual that allows one to control for competing factors. Randomized designs are broadly accepted as the most rigorous approach for constructing counterfactuals.

Randomized evaluations should be used much more often to combat IFFs. To the extent that randomized evaluations are infeasible, other methods that approximate the randomized ideal—quasi-experimental approaches—should be considered. A separate consideration is whether an intervention is sufficiently uniform across disparate cases so as to make appropriate comparisons, a task that is complicated when national or regional differences lead to qualitatively different interventions.

The task of evaluating the effects of global governance strategies on IFFs is further complicated if one looks beyond the flows themselves to various second-order outcomes.[15] Stemming IFFs is important in its own right, but frequently the goal is to reduce IFFs in order to address the predicate crimes linked to the flows (e.g., drug trafficking, corruption, terrorism). Alternatively, addressing IFFs is often directed toward broader macroeconomic outcomes such as development. If the goal is to evaluate these second-order outcomes, the task is substantially more difficult. Again, evaluating the effects on the first-order outcome of volume of IFFs is itself difficult. Indeed, as Levi and coauthors note, “if the right measure of AML success is a reduction in the volume of money laundering, there is little prospect of developing meaningful indicators at the national or global level.”[16]

Because randomization of a specific intervention (again, think about a global standard such as KYC) is not always or often possible, various other impact evaluation approaches seek to approximate a randomized design, though always with some compromises. Generally, quasi-experimental approaches begin from the premise that there exists a broken experiment to be fixed. In most cases, this means that the subjects under consideration could not be randomized to experimental conditions and so research design and econometric fixes need to be introduced to approximate randomization or produce comparisons of subjects that are as good as randomized. Of course, one can never reverse-engineer actual randomization, so the various quasi-experimental methods typically introduce some basic compromises. Even so, experimental and quasi-experimental designs hold substantially more promise than more basic case study comparisons or overall descriptive trends. The appendix provides a synopsis of the major impact evaluation approaches, including experimental, quasi-experimental, and observational approaches.

FORMAL AND INFORMAL COMPLIANCE WITH FINANCIAL ACTION TASK FORCE RECOMMENDATIONS

Concerned about the validity of current approaches to evaluating the FATF recommendations, Daniel Nielson, J. C. Sharman, and I carried out a global randomized experiment and associated audit study in the area of beneficial ownership and AML to measure whether FATF standards are effective. This is one of the few studies that attempt rigorous evaluation of global governance strategies to stem IFFs.[17]

Before carrying out the full randomized audit study, we established a baseline of formal compliance. Formal compliance refers to whether national governments enact laws that match international standards. Along with Shima Baradaran, we first culled statutory compliance levels from FATF's Mutual Evaluation Reports and set that as a baseline.[18] Because the enactment of a national law does not guarantee that corporate service providers (CSPs) in practice follow KYC rules to establish beneficial ownership, formal compliance measures cannot guarantee actual compliance.

A randomized experimental audit study on CSPs was carried out to measure informal compliance with FATF recommendations 10, 22, and 24 specifically.[19] Informal compliance refers to whether the organizations governed by the laws (e.g., CSPs tasked with following KYC rules) violate them. Using aliases, over the course of about two years, we approached nearly four thousand CSPs globally—each at least twice—and varied information about international standards as well as the risk associated with the approach to observe whether CSPs would comply with international standards and national statutes. In determining compliance, we considered whether CSPs required clients to provide required identity and residency documentation to set up a company.[20] In some conditions, we gave CSPs information about international standards, including penalties for noncompliance, and for U.S.-based CSPs information about enforcement of laws. We also varied the risk associated with the approach by posing as corrupt individuals or terrorist financiers. Importantly, CSPs were randomly assigned to different conditions to mitigate possible confounders.

The study offered a few important lessons. First, it compared formal and informal compliance measures, and demonstrated some vital differences, including that they only weakly correlate. Statutory compliance, as reported by FATF mutual evaluations, with recommendations 10 and 22 was not significantly related to the actual compliance rates found in the experiments, indicating that whether a country has adopted these recommendations has no apparent relationship with informal compliance.[21]

Second, receiving information about international standards, including penalties associated with noncompliance, did not change informal compliance levels relative to a placebo condition in which international standards were not invoked. Whether invoking FATF, the private Association of Anti–Money Laundering Specialists, penalties for not complying, or norms of appropriateness for complying, we did not observe appreciable change in actual compliance levels among CSPs.

Third, in the U.S. context, where a large number of service providers were compared across states specifically, in the likely event that U.S. states were systematically different from countries more generally, the study invoked the possibility that the Internal Revenue Service (IRS) would likely enforce penalties for noncompliance with international standards, something it has done aggressively in at least some cases. The IRS is the domestic agency that liaises with FATF and implements FATF standards in the country.[22] Informing CSPs in the United States that the IRS could take enforcement actions generated a statistically significant decrease in noncompliance, suggesting that at least some types of enforcement information motivate behavioral change. Given that national enforcement mechanisms are familiar to prospective violators, this result is both important and relevant to the broader discussion of the effectiveness of global governance strategies.

Fourth, international standards prescribe a risk-based approach for CSPs to screen potential customers. Under the risk-based approach, CSPs and financial institutions are supposed to screen customers to determine the risk of their being linked to corruption and terrorism. In this sense, international standards have been designed to address these predicate crimes, among others. Thus, CSPs are supposed to screen customers to prevent IFFs that can in turn facilitate these predicate crimes. In the study, although a corruption treatment did not alter compliance levels, signaling possible connections to terrorism did decrease noncompliance. This suggests that the risk-based approach to KYC rules are partially successful.

Finally, in contrast to conventional wisdom about tax havens, descriptively the study found that CSPs in tax havens were far more compliant than those in Organization for Economic Cooperation and Development (OECD) and developing countries, a result that is statistically significant. The long-standing international scrutiny directed toward tax havens has possibly led to greater levels of compliance there relative to other countries, though this is only a conjecture, as the experiment does not capture historical trends or their explanations. Despite these levels of compliance, tax havens could still attract significant money because companies seek to avoid taxes, even if they do not necessarily evade taxes. If this conjecture is correct, it suggests optimism about the possibility that global standards will lead to national standards that will be enforced at the locus of compliance among CSPs.

APPLYING EVALUATION METHODS TO MEASURES TO COMBAT ILLICIT FINANCIAL FLOWS

Evaluating measures to combat IFFs, as defined in the macroeconomic sense, is unlikely to yield valuable conclusions about specific policy recommendations for at least two reasons. First, different policies are likely to deter different kinds of IFFs (e.g., a customs reform might decrease trade misinvoicing, whereas a shell company reform might decrease flows of bulk cash) and be highly geography-dependent (e.g., a shell company reform might affect IFFs to and from Guatemala differently than those to and from Zambia). Second, different types of IFFs likely yield different cost-benefit calculations (e.g., inflows versus outflows, customs evasion versus capital controls evasion, drug money versus terrorism, etc.). If the hope is to evaluate IFFs from a broader macroeconomic perspective, then a pre-post or cross-sectional (when the application of standards varies) design may be all that is possible, but such a design will only produce tenuous conclusions about bundled interventions (i.e., the conglomeration of global and national policies).

Scholars and policymakers should consider targeted evaluations of specific policies applied to disaggregated categories of IFFs. Disaggregation is critical not only to avoid the measurement challenges but also because it enables the use of rigorous impact evaluation designs, especially randomized evaluations. As rigorous and targeted evaluations accumulate, scholars and policymakers should attempt to aggregate evaluations. If targeted evaluations of narrower IFF categories happen, the evaluation approach that produces the most rigorous counterfactual should be prioritized. As the extended example from our study demonstrates, it is possible to evaluate narrower IFF areas while maintaining a rigorous evaluation approach and global focus.[23] Given that the locus of compliance for international and national rules is primarily with CSPs and financial institutions, many more possibilities in this vein raise various directions for future research and policy evaluation.

Existing observational approaches, especially with regard to measuring trade misinvoicing, rely on strong assumptions that are unlikely to be overcome on their own. Governments could cooperate to carry out audit studies to generate more accurate measurements of trade misinvoicing or other indicators. In addition to measurement, randomizing certain strategies—say, at customs agencies—could enable audit studies that give precise estimates of causal effects. Possibilities include randomizing price or quantity of imports (or information about the imports) to customs officials, with country bilateral cooperation; and randomizing the ambiguity (or lack) of harmonized codes for comparison of the declarations at both borders.

While random assignment is a pivotal strength, we could not randomize international standards themselves. The study instead entailed randomization of information about the standards and risks, including on corruption and terrorism.[24] A critical question is whether—and which—other global or national strategies, or at least information about the global or national strategies, could be randomized to generate better counterfactuals for impact evaluation. One possibility is randomizing information about different FATF recommendations (e.g., politically exposed persons) to financial institutions.

Randomization will likely be impossible for evaluating many global or national strategies, but if isolating impact is important, perhaps other quasi-experimental strategies could be used to establish more appropriate counterfactuals. This raises a question of whether—and which—strategies would be amenable to quasi-experimental methods such as instrumental variables, regression discontinuities, or matching. Minimally, matching countries similar in many characteristics but different in their enforcement of global standards (such as country-by-country multinational corporation reporting or automatic exchange of tax information) could generate better counterfactual comparisons. Ideally, other methods such as regression discontinuities could be exploited, perhaps through the identification of thresholds that constrain the behavior of some financial institutions over others.

Although much attention has been given to measurement of IFFs, almost no work evaluates the effects of global governance strategies designed to curb IFFs. Progress in evaluation is unlikely to occur in the absence of a shift from the aggregate, macroeconomic level to consider the effects of global governance strategies on disaggregated financial flows.

ENDNOTES

[1] "Goal 16: Promote Just, Peaceful and Inclusive Societies," United Nations Sustainable Development Goals, http://un.org/sustainabledevelopment/peace-justice.

[2] Financial Action Task Force, *The FATF Recommendations: International Standards on Combating Money Laundering and the Financing of Terrorism and Proliferation* (Paris: FATF, 2012), http://www.fatf-afi.org/media/fatf/documents/recommendations/pdfs/FATF%20Recommendations%202012.pdf.

[3] The nine FATF-style regional bodies include the Asia/Pacific Group on Money Laundering, the Caribbean Financial Action Task Force, the Eurasian Group on Combating Money Laundering and Financing of Terrorism, the Eastern and Southern Africa Anti–Money Laundering Group, the Task Force on Money Laundering in Central Africa, the Financial Action Task Force of Latin America, the Intergovernmental Action Group Against Money Laundering in West Africa, the Middle East and North Africa Financial Action Task Force, and the Committee of Experts on the Evaluation of Anti–Money Measures.

[4] Richard K. Gordon, "The International Monetary Fund and the Regulation of Offshore Centers," in *Offshore Financial Centers and Regulatory Competition*, ed. Andrew P. Morriss (Washington, DC: American Enterprise Institute, 2010).

[5] Michael G. Findley, Daniel L. Nielson, and J. C. Sharman, *Global Shell Games: Experiments in Transnational Relations, Crime, and Terrorism* (Cambridge: Cambridge University Press, 2014); Peter Reuter, ed., *Draining Development? Controlling Flows of Illicit Funds from Developing Countries* (Washington, DC: World Bank, 2012), http://documents.worldbank.org/curated/en/305601468178737192/pdf/668150PUB0EPI0067848B09780821388693.pdf.

[6] Matthew Salomon and Joseph Spanjers, *Illicit Financial Flows to and From Developing Countries: 2005–2014* (Washington, DC: Global Financial Integrity, 2017), http://gfintegrity.org/wp-content/uploads/2017/05/GFI-IFF-Report-2017_final.pdf.

[7] Emile van der Does de Willebois et al., *The Puppet Masters: How the Corrupt Use Legal Structures to Hide Their Stolen Assets and What to Do About It* (Washington, DC: World Bank, 2011), http://star.worldbank.org/sites/star/files/puppetmastersv1.pdf; Financial Action Task Force, *Transparency and Beneficial Ownership* (Paris: FATF, 2014), http://www.fatf-gafi.org/media/fatf/documents/reports/Guidance-transparency-beneficial-ownership.pdf; Financial Action Task Force, *FATF Report to the G20: Beneficial Ownership* (Paris: FATF, 2016), http://www.fatf-gafi.org/media/fatf/documents/reports/G20-Beneficial-Ownership-Sept-2016.pdf.

[8] Salomon and Spanjers, *Illicit Financial Flows to and From Developing Countries*.

[9] Maya Forstater, "Illicit Financial Flows, Trade Misinvoicing, and Multinational Tax Avoidance: The Same or Different?," CGD Policy Paper 123, Center for Global Development, March 2018, http://cgdev.org/sites/default/files/illicit-financial-flows-trade-misinvoicing-and-multinational-tax-avoidance.pdf; Maya Forstater, "Defining and Measuring Illicit Financial Flows," in "Global Governance to Combat Illicit Financial Flows: Measurement, Evaluation, Innovation," Council on Foreign Relations, October 2018.

[10] Michael Levi, Peter Reuter, and Terrence Halliday, "Can the AML System Be Evaluated Without Better Data?," *Crime, Law and Social Change* 69, no. 2 (2017): 307–328, http://doi.org/10.1007/s10611-017-9757-4.

[11] Of course, there are a number of other challenging issues, such as whether IFFs should include legal flows (e.g., tax avoidance, capital flight) in addition to illegal flows, a topic that is beyond the scope of this paper.

[12] Levi, Reuter, and Halliday, "Can the AML System Be Evaluated Without Better Data?"

[13] See, for example, Raymond Fisman and Shang-Jin Wei, "The Smuggling of Art, and the Art of Smuggling: Uncovering the Illicit Trade in Cultural Property and Antiques," *American Economic Journal: Applied Economics* 1, no. 3 (2009): 82–96, http://doi.org/10.1257/app.1.3.82; Niels Johannesen and Gabriel Zucman, "The End of Bank Secrecy? An Evaluation of the G20 Tax Haven Crackdown," *American Economic Journal: Economic Policy* 6, no. 1 (2014): 65–91, http://doi.org/10.1257/pol.6.1.65; Jorgen Juel Andersen et al., "Petro Rents, Political Institutions, and Hidden Wealth: Evidence From Offshore Bank Accounts," *Journal of the European Economic Association* 15, no. 4 (2017): 818–860, http://doi.org/10.1093/jeea/jvw019; Annette Alstadsæter, Niels Johannesen, and Gabriel Zucman, "Who Owns the Wealth in Tax Havens? Macro Evidence and Implications for Global Inequality," *Journal of Public Economics* 162 (2018): 89–100, http://doi.org/10.1016/j.jpubeco.2018.01.008.

[14] J. C. Sharman, *The Money Laundry: Regulating Criminal Finance in the Global Economy* (Ithaca, NY: Cornell University Press, 2011).

[15] Peter Reuter and Edwin M. Truman, *Chasing Dirty Money: The Fight Against Money Laundering* (Washington, DC: Peterson Institute for International Economics, 2004); Miles Kahler, "Countering Illicit Financial Flows: Expanding Agenda, Fragmented Governance," in "Global Governance to Combat Illicit Financial Flows: Measurement, Evaluation, Innovation," Council on Foreign Relations, October 2018.

[16] Levi, Reuter, and Halliday, "Can the AML System Be Evaluated Without Better Data?"

[17] We carried out the study from 2010 to 2012 and published results of the entire study in Michael G. Findley, Daniel L. Nielson, and J. C. Sharman, *Global Shell Games: Experiments in Transnational Relations, Crime, and Terrorism* (Cambridge: Cambridge University Press, 2014). We also produced a shorter media summary, based on the first phase of the project, which served as the basis for a number of media reports including in the *Economist*, *New York Times*, NPR, TED, and *60 Minutes*, among others; see Global Shell Games (website), http://globalshellgames.com. We also produced a number of academic publications reporting on various dimensions of the study, such as compliance with international standards, whether international law really matters, what anonymous incorporation means for funding terrorism, as well as the methodology of evaluating international standards in this area. See Michael G. Findley, Daniel L. Nielson, and J. C. Sharman, "Causes of Noncompliance With International Law: A Field Experiment on Anonymous Incorporation," *American Journal of Political Science* 59, no. 1 (2015): 146–161, http://doi.org/10.1111/ajps.12141; Michael Findley, Daniel Nielson, and J. C. Sharman, "Orchestrating the Fight Against Anonymous Incorporation: A Field Experiment," in *International Organizations as Orchestrators*, eds. Kenneth W. Abbott, Phillip Genschel, Duncan Snidal, and Bernhard Zangl (Cambridge: Cambridge University Press, 2015); Shima Baradaran et al., "Does International Law Matter?," *Minnesota Law Review* 97, no. 3 (2013): 743–837, http://doi.org/10.2139/ssrn.2097852; Shima Baradaran et al., "Funding Terror," *University of Pennsylvania Law Review* 162, no. 3 (2014): 477–536, http://jstor.org/stable/pdf/24247862.pdf; Michael G. Findley et al., "External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation," *Journal of Politics* 79, no. 3 (2017): 856–872, http://doi.org/10.1086/690615; Brent B. Allred et al., "Anonymous Shell Companies: A Global Audit Study and Field Experiment in 176 Countries," *Journal of International Business Studies* 48, no. 5 (2017): 596–619, http://doi.org/10.1057/s41267-016-0047-7; Michael G. Findley, Daniel L. Nielson, and J. C. Sharman, "Using Field Experiments in International Relations: A Randomized Study of Anonymous Incorporation," *International Organization* 67, no. 4 (2013): 657–693, http://doi.org/10.1017/S0020818313000271.

[18] Baradaran et al., "Does International Law Matter?"

[19] Recommendation 24 is most applicable to the case of shell companies. It states: "Countries should take measures to prevent the misuse of legal persons for money laundering or terrorist financing. Countries should ensure that there is adequate, accurate, and timely information on the beneficial ownership and control of legal persons that can be obtained or accessed in a timely fashion by competent authorities." Other relevant sections of the Financial Action Task Force recommendations mandate that corporate service providers follow careful in-depth customer due diligence and record-keeping responsibilities when establishing business relationships (recommendation 22). Specifically, CSPs are supposed to "[identify] the customer and [verify] that customer's identity using reliable, independent source documents, data, or information" (recommendation 10), later articulated to be passports, national identity cards, or drivers' licenses. See Financial Action Task Force, *The FATF Recommendations*.

[20] We only considered the first-order question of whether CSPs would offer anonymous incorporation. An important follow-on question, which would be interesting to investigate, would be about the effects of these anonymous incorporations, a question that will have to await further examination.

[21] Statutory compliance with recommendation 24 was weakly related to actual compliance, but only at the 0.1 threshold (p = 0.08), and the correlation was weak (r = 0.26). A regression analysis indicates that 93 percent of the variance in actual compliance rates remains unexplained by statutory compliance levels.

[22] "Criminal Investigation Responds to Terrorism," Internal Revenue Service, accessed December 15, 2012, http://web.archive.org/web/20170517025326/irs.gov/uac/Criminal-Investigation-Responds-to-Terrorism; "International Investigations: Criminal Investigation (CI)," Internal Revenue Service, accessed December 15, 2012, http://irs.gov/compliance/criminal-investigation/international-investigations-criminal-investigation-ci.

[23] Findley, Nielson, and Sharman, *Global Shell Games*.

[24] Findley, Nielson, and Sharman, *Global Shell Games*.

Appendix: Evaluation Strategies for Global and National Measures Against Illicit Financial Flows

The construction of a rigorous counterfactual is the most fundamental consideration for evaluating the impact of global governance strategies.* That is, for a given global governance strategy (e.g., anti–money laundering policy), it is critical that some subjects of study receive the policy treatment whereas others do not, and that both of those sets of subjects otherwise be identical (or highly similar) in every respect. Only then, through explicit comparison with subjects that did not receive the policy treatment, can an evaluation establish that those that did receive the treatment changed their behavior accordingly.

A rigorous counterfactual can be constructed in several ways, which are broadly categorized as experimental, quasi-experimental, and observational approaches. The most credible are experimental and quasi-experimental approaches. In the discussion below, global governance strategies are referred to as programs or treatments that one seeks to evaluate. Given the goal to attribute any differences in outcomes to the program and not to other factors, these other possibilities are referred to as potential confounders.

EXPERIMENTS

The defining feature of an experiment is that subjects are randomly assigned to treatment or control conditions. Randomization to treatment and control is often considered the ideal approach to identify causal impact because random assignment makes the control and treatment groups' characteristics, in expectation, identical. In other words, through random assignment, the control group constitutes a rigorous counterfactual because it is theoretically identical to the treatment group, except that the treatment group receives an intervention. Any difference in outcomes between the two groups can only be attributed to the intervention, given that the groups are identical in every other respect. In this respect, experiments have high internal validity and can offer the most credible answer to the question of whether global governance strategies work. To use randomization, a few prerequisites must be satisfied: it must be possible to give the treatment to some entities, but not others; the treatment needs to be uniform or consistent in its application; there must be a sufficient number of entities to enable balancing of possible confounding factors; and threats to validity, such as attrition, noncompliance, and interference, must be preventable. Michael Findley and the coauthors provide

* This appendix surveys major impact evaluation methods and examines their strengths and weaknesses. For detailed and technical discussions of these various impact evaluation approaches, see Alan Gerber and Donald Green, *Field Experiments: Design, Analysis, and Interpretation* (New York: W. W. Norton, 2012); Rachel Glennerster and Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide* (Princeton, NJ: Princeton University Press, 2013); and Joshua D. Angrist and Jorn-Steffen Pischke, *Mastering Metrics: The Path From Cause to Effect* (Princeton, NJ: Princeton University Press, 2014).

one of the few examples of randomized evaluations in the research of illicit financial flows.[†] Of course, it is not always possible to carry out a randomized study, and therefore quasi-experimental strategies need to be employed.

QUASI-EXPERIMENTS

In the event randomization is not possible, other methods to produce a rigorous counterfactual need to be considered. When necessary, impact evaluators typically consider several quasi-experimental approaches: regression discontinuity designs (RDDs), difference in differences, instrumental variables, and matching.

- *Regression discontinuity designs:* RDDs can be used to evaluate programs with arbitrary and strictly enforced eligibility cutoffs. Typical implementations include selection into a program based on, for example, income status just above or below a threshold. From a causal identification perspective, it is extremely important that participants and nonparticipants—who are just above and below the cutoffs—are identical in every respect except that some are assigned to the program whereas others are not. As such, any differences in outcomes should be attributable to the program and nothing else. RDDs can be especially useful because they balance both observable and unobservable potential confounders. However, it is often difficult to identify programs that are implemented based on the arbitrary and strictly enforced eligibility criteria. Unlike randomized evaluations, RDDs can be used for ex-post evaluation, as long as sufficient data exists.

- *Difference in differences:* This design couples a before-after comparison of program participants with a cross-sectional design of participants and nonparticipants. In this case, changes in outcomes over time for program participants can be compared to changes in outcomes over time for nonparticipants. If relevant assumptions are satisfied, difference-in-differences analyses can account for observable and unobservable potential confounders. The most difficult assumption to satisfy, however, is that of parallel trajectories. That is, had the program not existed, the two groups—program participants and nonparticipants—would have had identical trajectories over the period in question, an assumption that is difficult to satisfy in practice.

- *Instrumental variables:* This approach is often used after a program is implemented and seeks to separate possible confounding information from unique program information that can be referred to as plausibly exogenous program effects. An instrumental variables approach works by identifying a third variable that is highly correlated with the program but uncorrelated with other factors that could affect the outcome of interest. In an intent to treat analysis, the original randomization is used as an instrument for program uptake. Otherwise, some other third variable could be used. In either case, if satisfied, the instrumental variable approach helps separate the unique program effects from confounders. If a suitable instrument can be found—a difficult task in most cases—then observable and unobservable potential confounders can be ruled out in reaching conclusions about the effect of the program.

[†] Michael G. Findley, Daniel L. Nielson, and J. C. Sharman, *Global Shell Games: Experiments in Transnational Relations, Crime, and Terrorism* (Cambridge: Cambridge University Press, 2014).

- *Matching:* This approach uses observable characteristics to create matches of participants and nonparticipants. That is, using statistical algorithms, one creates matched pairs in which the entities in the pairing are identical (or highly similar) in all respects except that some are program participants whereas others are not. If such balance can be created through matching, then any differential outcome should be attributable to program status. A difficult challenge with matching approaches is that any potential confounders that are unobservable cannot be accounted for unless one can include observable factors that are plausibly correlated with the critical unobservable factors.

OBSERVATIONAL DESIGNS

It is not always possible to produce a rigorous counterfactual through experimental or quasi-experimental methods. In such cases, impact evaluations sometimes employ standard regression-based statistical analyses, time series analyses such as pre-post comparisons, cross-sectional comparisons such as participant-nonparticipant comparisons, or qualitative approaches such as process tracing.

- *Multiple regression:* In a regression framework, program participants and nonparticipants are compared while controlling for other factors that could explain the differences. The pivotal assumption is that control variables included in the model capture all relevant ways in which the two groups of subjects may differ. In other words, one must ensure that all characteristics that could be correlated with outcomes—both observable and unobservable—are captured in the regression. Of course, unobservable characteristics cannot be included in a regression analysis and therefore cannot be ruled out with any confidence.

- *Pre-post comparisons:* In such a design, evaluators compare outcomes for participants both before and after a program has been implemented. In this case, the comparison group includes the participants themselves before the program was implemented. The pivotal assumption here is that the program was the only factor influencing changes in the measured outcomes over time. Unfortunately, many factors can change concurrently with the program and can affect the outcomes for the participants, and yet it is extremely difficult to separate the effects of the program from those of potential confounders.

- *Participant-nonparticipant comparisons:* In this design, the evaluation relies on comparing the outcomes of program participants and nonparticipants only after the program is implemented. The pivotal assumption is that participants and nonparticipants are identical, especially in that they are equally likely to enter the program. Unfortunately, many reasons explain why some are selected (or select) into programs whereas others are not (or do not), including motivation to take up the program or fitting the demographic of needing the program. Preexisting differences between participants and nonparticipants are likely more responsible for differences in outcomes than the program itself.

- *Qualitative approaches:* Sometimes quantitative and qualitative strategies are pitted against each other, but this is likely a false dichotomy. Both should be used in tandem, wherever appropriate. Interviews and focus groups are critical for getting at mechanisms, exploring ideas that evaluators had not thought of, and allowing interactions across respondents through focus groups. By themselves, such approaches cannot produce credible inferences, but they can help add important context and details otherwise missing from other impact evaluation methods.