

Impact Evaluation of Trade Interventions

Paving the Way

Olivier Cadot
Ana M. Fernandes
Julien Gourdon
Aaditya Mattoo

The World Bank
Development Research Group
Trade and Integration Team
November 2011



Abstract

The focus of trade policy has shifted in recent years from economy-wide reductions in tariffs and trade restrictions toward targeted interventions to facilitate trade and promote exports. Most of these latter interventions are based on the new mantra of “aid-for-trade” rather than on hard evidence on what works and what does not. On the one hand, rigorous impact-evaluation is needed to justify these interventions and to improve their design. On the other hand, rigorous evaluation is feasible because unlike traditional trade policy, these interventions tend to be targeted and so it is possible to construct treatment and control groups. When interventions are not targeted,

such as in the case of customs reforms, some techniques, such as randomized control trials, may not be feasible but meaningful evaluation may still be possible. This paper discusses examples of impact evaluations using a range of methods (experimental and non-experimental), highlighting the particular issues and caveats arising in a trade context, and the valuable lessons that are already being learned. The authors argue that systematically building impact evaluation into trade projects could lead to better policy design and a more credible case for “aid-for-trade.”

This paper is a product of the Trade and Integration Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The authors may be contacted at Olivier.Cadot@unil.ch, afernandes@worldbank.org, julien.gourdon@cepii.fr, amatoo@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Impact Evaluation of Trade Interventions: Paving the Way¹

Olivier Cadot^{*}
Ana M. Fernandes⁺
Julien Gourdon[§]
Aaditya Mattoo⁺⁺

JEL classification codes: F13, F14, L15, L25, O17, O24, C23

Keywords: impact evaluation, trade competitiveness, trade facilitation, aid for trade, export promotion, randomized control trials, propensity-score matching.

¹ We thank Vivian Agbegha for excellent research assistance, Christina Neagu for help with tariff data and Mohini Datt for help with data on World Bank aid for trade. We thank Daniel Lederman, Martin Ravallion, and participants at the December 2010 workshop on “Impact Evaluation of Trade Interventions: Paving the Way” in Washington, DC for comments. Support from the governments of Norway, Sweden and the United Kingdom through the Multi-Donor Trust Fund for Trade and Development is gratefully acknowledged. This paper is the result of collaboration between the World Bank and Switzerland’s NCCR on the evaluation of trade-related interventions.

^{*} University of Lausanne, CEPREMAP and CEPR.

⁺ Trade and International Integration Unit, Development Economics Research Group, World Bank.

[§] CEPII.

⁺⁺ Trade and International Integration Unit, Development Economics Research Group, World Bank.

1. Introduction

Trade policy has changed fundamentally since the days of structural adjustment and economy-wide trade reforms. Partly in reaction to the uneven results of trade policy reforms, the focus has shifted to more targeted interventions aimed at reducing trade costs and addressing market failures that inhibit exports. Significant national resources and international assistance are now devoted to trade facilitation and export promotion, and the international development community has galvanized around a new “aid-for-trade” (AFT) mantra as a means of helping low-income countries integrate into the global economy.

The environment in which trade-related assistance is provided has also changed. In times of fiscal austerity, taxpayers increasingly question the justification for large aid flows and, at the very least, demand results and accountability.² The development community has struggled to respond to these demands because there is surprisingly little evidence about what works and what doesn't in the area of trade and industrial policies.

An authoritative survey of trade and industrial policy recently acknowledged that there is hardly any microeconomic evidence to guide specific trade interventions (Harrison and Rodríguez-Clare, 2010). There are several reasons for the disappointing pace at which such evidence has been gathered. Trade policy research has been slow to respond to changing needs. Tariffs continue to occupy center stage in policy research, in spite of their declining importance as trade barriers, simply because they are easy to measure. The aid-for-trade community has in turn been slow to build a culture of rigorous evaluation. For instance, a review of 85 recent World Bank trade-related projects conducted by the authors revealed that only five of them included rigorous evaluation components. Worse, those few evaluations relied on crude before-after comparisons, which are known to be vulnerable to confounding influences. The “knowledge-market failures” identified by Ravallion (2009) have also inhibited rigorous evaluations in the trade context: demanders of knowledge about the effectiveness of trade interventions have inadequate information about the quality of any potential evaluation, especially because there are so few good examples; project managers tend to have “monopolistic” control over which projects get evaluated, at what cost and how; and the benefits from the rigorous evaluation of a particular trade project accrue in large part to other future projects which do not share in the cost of evaluating the project.

Still, trade evaluation itself can benefit from the positive externalities generated by research in other areas. In fact, the tools for a serious evaluation of trade-related interventions are already

² A recent poll featured by the Financial Times (Financial Times, July 12, 2010) showed that a majority of respondents in OECD countries considered defense and development aid as priority areas for spending cuts.

there. Originally developed in the agro-biological and then the medical sciences, impact evaluation (IE) methods have spread to the social sciences and are routinely employed in the areas of health and education. In essence, an impact evaluation compares the outcomes of entities — individuals or firms — that received support from a program or were directly impacted by a policy with the counterfactual outcome of those same entities had the program or policy not been in place. Because such counterfactual outcomes are not observable, they are approximated by the outcomes of a control group.

IE methods have provided powerful tools in other fields to help guide policy choices and minimize the cost of interventions. For instance, Banerjee and Duflo (2008) showed how a comparison of IE results established that, in order to raise school attendance rates among Kenyan children, a program to treat intestinal worms was twenty times more cost-effective than hiring teachers, suggesting a clear prioritization of actions.³

The recent creation by the World Bank of a separate impact evaluation unit as part of the Development Impact Evaluation Initiative (DIME) has helped spread IE methods to new areas of development research and practice.⁴ For instance, World Bank researchers have led the way in analyzing the impact of business registration reform or bankruptcy reform (Klapper and Love, 2010; Bruhn, 2011; Gine and Love, 2011). Researchers have also begun to use these methods to evaluate programs and policies in the area of private sector development, where the treated "entities" are firms (see McKenzie, 2010 for a survey). Similar evaluations could be used to guide trade interventions.

The usual excuse for *not* using IE methods in assessing the effectiveness of trade assistance is that the "clinical" nature of the treatment needed for a proper definition of treatment and control groups is absent from trade policy. This was perhaps true of old-style trade policies like structural adjustment or tariff reforms; but it is not true of the new trade interventions like export promotion. This paper intends to show that *trade exceptionalism* — the notion that trade-related interventions are inherently not amenable to IE — is, if anything, limited to traditional trade policies. More recent, focused trade-related interventions can be evaluated formally, provided that one is not wedded to a particular methodology such as randomized-control trials (RCTs). Although, as we will see, the range of application of RCTs is broader than one might think, other quasi-experimental methods are available and can shed light on what works and what does not.

³ This ratio was established by comparing the evaluation of a de-worming program by Miguel and Kremer (2004) with a separate evaluation of a program to reduce teacher-student ratios by Banerjee, Jacob, Kremer, Lanjouw, and Lanjouw (2005). Comparing impact estimates from separate impact evaluations is tricky since each has been established in a particular context with limited external validity (we will return to the issue later on in this paper). However, when the difference in cost effectiveness is as large as this one, the risk of getting the prioritization order wrong is reduced.

⁴ Information on DIME can be obtained at: <http://web.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDEVIMPEVAINI/0,,menuPK:3998281~pagePK:64168427~piPK:64168435~theSitePK:3998212,00.html>.

RCTs are only one of the possible approaches for rigorous impact evaluation. For instance, some countries implement regulatory reforms in staggered fashion starting in a small set of locations before extending them to all locations. The impact of such reforms can be rigorously evaluated by using locations where the reforms are introduced later as a control group for the locations where reforms are introduced earlier and using a difference-in-differences estimation methodology (see Bruhn, 2011). Similarly, ex-post evaluation of programs and policies is a possible approach, provided that information is available both on which firms received support from a program or were directly impacted by a policy as well as on the entire (or a large portion of the) universe of firms. In these circumstances, it is possible to use propensity score matching combined with difference-in-differences estimation (see e.g. Tan, 2009; Lopez-Acevedo and Tinajero, 2010).

These methods have already been applied in a number of recent studies and have produced interesting and unexpected results. Consider the following three examples:

First, in an ex-post evaluation of export promotion programs in six Latin American countries using rich firm-level datasets, Volpe (2011) shows that these programs were effective in facilitating export expansion primarily along the extensive margin (i.e., through an increase in the number of products exported or in the number of export markets served) rather than along the intensive margin (an increase in exports of existing products to existing markets). He also shows that programs benefitted small and relatively inexperienced firms more than larger and already established exporters, and that bundled services providing support to firms throughout the export development process were more effective than isolated actions.

Gourdon, Marchat, Sharma, and Vishwanath (2011) use similar ex-post evaluation methods to assess the impact of a World Bank-financed export promotion program in Tunisia — FAMEX — which provided a mixture of counseling and matching grants to new exporters. Their findings suggest that export promotion has a large and significant effect on overall export growth: a 39% increase in the average annual growth rate of program beneficiaries relative to the control group over a four-year period. The effect of the program on the extensive margin of exports – in terms of products and destinations – is more subdued: about 5% higher growth for beneficiaries that is significant only for destinations. They also find a significant increase in employment growth, i.e., 10% more for program beneficiaries than for control firms. The effect on export growth is stronger for firms that were initially only marginal exporters (exports represented less than 20% of turnover). Interestingly, their sample also includes services firms, for which the effect of export promotion is significantly larger than for manufacturing firms.

Datt and Yang (2011) analyze a natural experiment in which the Philippines government suddenly reduced the minimum value threshold under which shipments were exempt from pre-shipment inspections (PSI), closing a loophole that had encouraged importers to slice shipments

in order to escape inspection. They show that the reform failed to curb under-invoicing and thus to raise duty collection as importers switched to an alternative loophole, namely, the use of an export-processing zone (EPZ). As this alternative loophole involved high fixed costs (setting up a presence in the EPZ), in the end the Philippine government was no better off while importers were worse off. The authors also discuss the effects of a related policy reform in Colombia where the government sought to remedy undervaluation of certain imports by mandating PSI on a subset of products. This, however, left open the loophole of misclassification of those products as similar products that did not require a PSI. Both cases illustrate the importance of careful, incentive-compatible reform design.

This paper considers a detailed menu of trade-related interventions and discusses the challenges posed by their evaluation. In doing so, we discuss examples of impact evaluations using a range of methods (experimental and non-experimental) highlighting the particular issues and caveats arising in a trade context, and the valuable lessons that are already being learnt. We argue that systematically building impact evaluation into trade projects could lead to better policy design and to a more credible case for “aid-for-trade.” The rest of the paper is organized as follows: Section 2 discusses the changing nature of trade policy while Section 3 reviews the available evidence on the impact of trade assistance. Section 4 considers trade-related interventions and their evaluation. Section 5 addresses the data issues crucial to impact evaluation. Section 6 discusses the future challenges in IE of trade assistance. Section 7 concludes.

2. The changing nature of trade policy and trade assistance

Most developing countries have moved beyond the first generation of trade reforms, which involved across-the-board cuts in tariffs and the elimination of import quotas. Tariffs have fallen substantially over the last 20 years. The simple average applied tariff of World Trade Organization (WTO) members on all goods was 5.8 percent in 2008 (WTO, 2009), and the developing country average is down to around 10 percent compared to 30 percent in 1990.

Recourse to quantitative restrictions has also substantially declined. One reason is the narrower interpretation of the *balance-of-payments exception* in the WTO and the stricter enforcement of the conditions under which it can be invoked. Countries like India have been forced to phase out numerous quotas that had been maintained for a long time, ostensibly to address balance of payments difficulties. Another reason is the tighter interpretation following the Uruguay Round Agreement of the *national treatment* provision in the WTO, which precludes local-content requirements that many developing countries had favored and other members had tolerated.

With this decline in traditional barriers to market access, supply-side constraints are seen as the main obstacle that developing countries face in taking advantage of new opportunities in

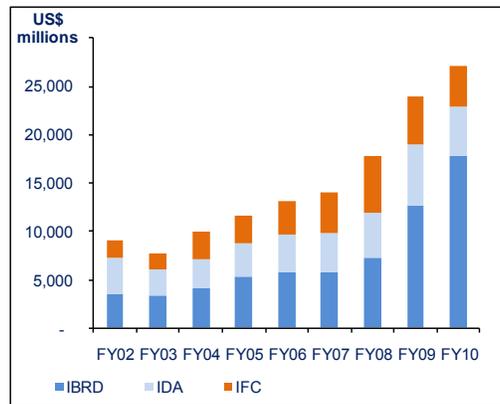
international markets. Therefore, trade interventions are becoming more targeted, focusing either on (a) the *trade facilitation* agenda, involving, for example, customs reforms and infrastructure — e.g. port — improvements and/or (b) the *trade competitiveness* agenda, consisting of proactive industrial policies, involving productive capacity building, EPZs, or export promotion. In designing such trade interventions, developing countries need policy advice, particularly more evidence-based advice. They need to know which interventions work and which do not, in which sectors, in which sequence, and which ones are most cost-effective.

The World Bank too has shifted emphasis in its trade assistance from broad trade liberalization reforms in the 1980s and 1990s to more targeted interventions to reduce the costs of trade and to equip producers to export since the early 2000s. The declaration of World Trade Organization (WTO) ministers in Hong Kong SAR, China in 2005 and the first Global Aid for Trade Review in Geneva in 2007 gave an impetus to the expansion of aid for trade to help developing countries build their supply-side capacity and trade-related infrastructure.

The World Bank responded by expanding its commitments on trade competitiveness, trade facilitation, and infrastructure, and is now a leading contributor to aid for trade. As shown by Figure 1, recent commitments by the World Bank are substantial and growing: concessional trade-related lending (as per OECD/WTO definition) to low-income countries grew from US\$3.18 billion annually in 2002–2005 to an average of US\$4.84 billion in 2007–2008, while non-concessional trade-related lending to middle income countries increased from US\$4.16 billion in 2002–2005 to US\$9.8 billion in 2007–2008 (World Bank, 2011).⁵ Since 2001 the World Bank approved 437 trade-related lending projects in 90 countries and 53 trade-related lending operations in 10 regional groups, with Africa and Eastern Europe and Central Asia accounting for most of the operations (World Bank, 2011).

Figure 1
World Bank aid-for-trade commitments 2002–2010

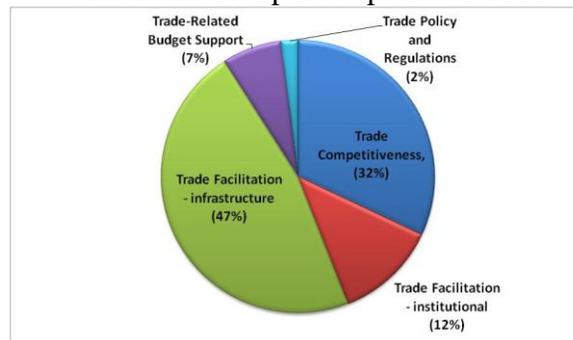
⁵ The numbers presented in the figure are based on the OECD/WTO definition of aid-for-trade. The sectors that fall under this definition are (1) for IBRD/IDA - agriculture, fishing and forestry; information and communication; energy and mining; transportation; and industry and trade; (2) for IFC - agriculture and forestry; information; oil, gas and mining; chemical; utilities; transportation and warehousing; construction and real estate; food and beverages; nonmetallic mineral product manufacturing; primary metals; pulp and paper; textiles, apparel and leather; plastics and rubber; industrial and consumer products; wholesale and retail trade; professional, scientific and technical services; and accommodation and tourism services.



Source: Authors' calculations based on data from the World Bank Business Warehouse website.

Trade facilitation-related infrastructure is the largest single component of World Bank trade-related investments in developing countries, while the rest consist mostly of improving competitiveness. Figure 2 shows the distribution of World Bank commitments on aid for trade as of fiscal year 2008 while Table 1 details the types of interventions falling under the "trade competitiveness" and the "trade facilitation" agendas.

Figure 2
World Bank Group trade portfolio 2008



Source: Authors' calculations based on data from the World Bank Business Warehouse website.

Given the increase in aid for trade, donors and recipients would like to see evidence that this new type of assistance will be more effective than past aid efforts. These concerns are especially strong in the aftermath of the recent financial crisis, when pressures to reduce fiscal deficits and debt are weakening political support for foreign assistance. In fact, a recent opinion poll in OECD countries revealed that a large majority of the public favored cuts in defense and aid spending rather than in other categories of expenditure.⁶

Table 1
Focused trade interventions

<i>Trade competitiveness (including trade finance)</i>	<i>Trade facilitation and logistics</i>
--	---

⁶ See *Financial Times* of July 12, 2010.

<input type="checkbox"/> Export promotion/diversification <input type="checkbox"/> Support to producer/exporter organizations <input type="checkbox"/> Quality testing and export certification <input type="checkbox"/> Technology upgrading and support services <input type="checkbox"/> Strengthening policy/regulatory framework <input type="checkbox"/> Export credit insurance <input type="checkbox"/> Export credit guarantee <input type="checkbox"/> Line of credit <input type="checkbox"/> Support for financial institutions	<input type="checkbox"/> Customs reform <input type="checkbox"/> Ports/airports rehabilitation <input type="checkbox"/> Railway privatization/rehabilitation <input type="checkbox"/> Roads construction/rehabilitation
---	--

3. Existing methods of evaluating trade interventions

In this section, we review three kinds of existing evaluation efforts. The first involves broad and—to this day—largely inconclusive assessments of aid for trade and its impact. The second examines the effect of national trade interventions, such as export promotion activities, but still at a highly aggregate level, considering mostly aggregate exports as outcomes; this literature provides some support for certain types of focused interventions. The third set of efforts involves assessments by the World Bank of its own trade-related projects. While this last set are in principle as focused as the interventions themselves, they have for the most part not been based on the collection or analysis of any hard evidence on impact.

Before we discuss some of the results emerging from each strand of evaluation efforts, it is worth noting that cross-country regressions, on which the first two strands rely heavily, have strengths of their own, but also limitations that are sufficiently serious to have prompted a growing number of development scholars to turn to different methodologies if not an altogether different paradigm. On the positive side, cross-country regressions—based on either cross-sections *stricto sensu* or on multi-period panels—provide general average estimates of the effects of a policy or program that are not reflections of a specific context. They also pick up the entire effect of the policy or program, including externalities and general-equilibrium feedbacks. Both of these strengths are particularly relevant in comparison with micro-level impact evaluations, as we will see later on.

On the negative side, like the earlier literature on the effect of trade reforms, cross-country regressions evaluating the effect of aid or specific trade interventions tend to suffer from problems of weak identification and attribution.⁷ Neither policies nor aid flows can be taken as exogenous to the performance outcomes they are supposed to affect, and no instrumental-

⁷ For a thorough discussion of the trade-off between internal and external validity, see for instance Rodrik (2008) and references therein.

variable strategy, however clever, has dispelled doubts about reverse causation or omitted-variable bias, both likely to be present at the level of aggregation at which these studies are cast. Impact evaluations, for all their own limitations, are less vulnerable to these identification issues, because they rely, for identification, on outcome differences between treatment and control groups *in the same context*, instead of variations in policy choices or aid flows across countries.

3.1 Evaluating aid for trade

The literature on the impact of AfT is fairly limited, in part because AfT projects are not always distinguishable from other aid projects. As in the rest of the aid effectiveness literature, the results are ambiguous (Rajan and Subramanian, 2008). Regarding the cross-country allocation of AfT, Gamberoni and Newfarmer (2009) find that, after controlling for absorption capacity (related, for example, to governance), more AfT is directed towards countries with a higher demand for AfT as measured by indicators of “underperformance” in trade.⁸

On impact, one strand of the literature explores whether AfT positively affects exports from the donor country to the recipient country given that, up to the early 1990s, over half of all bilateral aid was at least partially tied to donor exports. Using a gravity equation, Wagner (2003) shows that this form of trade was indeed boosted; but Osei, Morrissey, and Lloyd (2004), using a gravity equation in first differences for a panel of four European donors and 26 African recipients, found an unstable and insignificant impact of aid on exports from donor to recipient. Recently, Nelson and Juhasz Silva (2008) use a more conventional gravity equation including bilateral aid flows as a regressor (instrumented by their one-year lagged value), and find a significant although small impact on trade flows from donor to recipient.

From a development perspective, only a few of the recent studies focus on the more relevant question of whether aid raises the export capacity of recipient countries. Cali and te Velde (2011) regress trading costs and the value of exports on lagged AfT disbursements and control variables, using data from the OECD's Creditor Reporting System that separately identifies aid to trade facilitation and infrastructure from aid to productive capacity.⁹ Using a large panel of developing countries, the authors address the possibility of endogeneity and measurement errors in AfT flows by instrumenting those with the Freedom House's index of civil liberties. The message that emerges across their various specifications is that aid to trade facilitation and infrastructure seems to have a significant effect in reducing trade costs and in increasing export values, while

⁸ Underperformance in trade is captured by multiple indicators. Countries that underperform in trade can be those in the lower two quintiles of performance measured along five dimensions: (a) those experiencing relatively slow growth of exports of goods and services, (b) those losing global market share, (c) those suffering deterioration in competitiveness in existing markets, (d) those exporting slow-growing products or to slow-growing markets, and/or (e) those over-reliant on only a few exports. Also, countries that underperform in trade are those that under-trade with bilateral partners, controlling for market size and distance, those with low levels scores on the World Bank logistics performance index for transport or for customs, and on an indicator of peak tariffs.

⁹ Trading costs are measured by the trading across borders indicators of the Doing Business database.

aid to productive capacity is insignificant. When considering sectorally targeted aid, the authors again find that aid to infrastructure has a significant impact on export values, but aid to productive capacity does not, controlling for country-sector fixed effects that account for comparative advantage differences.

Brenton and von Uexkull (2009) examine the response of product-level exports from developing countries to product-level export-development aid, combining mirrored product-level (HS4) export data with export-development aid data from the German cooperation agency GTZ and from the OECD/WTO Trade Capacity Building Database for 48 developing countries. Using a matching difference-in-differences (DID) approach (discussed in section 4) they show insignificant effects of contemporaneous and lagged aid on product-level exports after controlling for lagged exports, and country and year-product fixed effects, and eliminating outliers.¹⁰ However, the authors do show strong positive effects in a simple comparison of product-level exports before and after receiving export development aid. This finding suggests an important attribution problem — namely, export growth may not be due to the aid received but instead may reflect the fact that aid targets sectors with promising prospects. The authors go on to argue that, in evaluating the impact of technical assistance for exports, it is essential to identify what would have happened in the absence of the policy intervention. This is a primary concern in this paper.

As the literature stands, it is fair to say that the effect of AfT on the export performance of beneficiary countries has not been established on the basis of aggregate numbers. Ferro, Portugal-Perez, and Wilson (2011) advance the analysis of the effectiveness of AfT revisiting the data from OECD's Creditor Reporting System. The authors exploit the differential intensities of service use across manufacturing sectors (based on input-output tables from the U.S. and Argentina) to evaluate the impact of aid for trade flows directed at five services sectors — transport, communications, energy, banking/financial services, and business services — on the exports of downstream manufacturing sectors in 106 aid-recipient countries over the period 1990–2008. Their identification strategy aims at circumventing reverse causality problems common in the AfT literature; and their results show that aid flows directed at the energy and banking sectors have a significant positive impact on downstream manufacturing exports.

3.2 Evaluating national trade interventions

A few recent cross-country studies suggest a positive impact of certain types of trade interventions, regardless of whether they are financed by donors or domestic government budgets. On export promotion, Lederman, Olarreaga, and Payton (2010) examine the

¹⁰ Their matching approach pairs each treatment country that receives export development aid for a given product i to the country that is more similar to it in terms of its likelihood to export product i , where this likelihood is estimated based on observable country characteristics such as the level of development, factor endowments, and climate conditions.

effectiveness of export promotion agencies (EPAs) based on a rich survey of EPAs across 88 developed and developing countries. The goals of EPAs are to help exporters understand and find markets for their products and services and can be divided into four categories: (a) country image building (advertising, promotional events, but also advocacy); (b) export support services (exporter training, technical assistance, capacity building, including regulatory compliance, information on trade finance, logistics, customs, packaging, pricing); (c) marketing (trade fairs, exporter and importer missions, follow-up services offered by representatives abroad); and (d) market research and publications (general, sector, and firm-level information, such as market surveys, on-line information on export markets, publications encouraging firms to export, importer and exporter contact databases) (Lederman et al. [2010], pp. 257–258). For 21 of the 73 developing countries surveyed, the authors find that EPAs receive budgetary support from multilateral donors such as the World Bank. The authors estimate the effect of EPAs' expenditures per capita on overall exports per capita at the country level, accounting for selection bias in survey responses and for potential reverse causality. Their main conclusion is that, on average, EPAs have a significant positive effect on exports. Their estimates also point to the importance of EPAs' services for overcoming foreign trade barriers and solving asymmetric information problems associated with exports of differentiated goods. In addition, they find evidence of strong diminishing returns, suggesting that small is beautiful as far as EPAs are concerned. However, the authors acknowledge that cross-country regressions cannot fully capture the heterogeneity of policy environments and institutional structures in which EPAs operate; hence, more detailed studies or project-type analyses are needed to provide specific policy advice.

On trade facilitation, Helble, Mann, and Wilson (2009) examine the responsiveness of trade flows to various types of aid for trade — linked to reform of trade policy and regulation, trade development (productive capacity building), and economic infrastructure — using a gravity equation framework covering 167 importers (reporters) and 172 exporters (partners) during the 1990–2005 period. Their results indicate that relatively small amounts of aid targeted at trade policy and regulatory reform have a greater impact with respect to increased trade flows than aid for broad trade development assistance or infrastructure. Several recent papers point to the importance of internal barriers related to infrastructure and institutions — including logistics performance — as obstacles to developing countries' ability to trade and the volume of trade (e.g., Djankov, Freund, and Pham, 2010; Francois and Manchin, 2007; Freund and Rocha, 2011; Hoekman and Nicita, 2008; Portugal-Perez and Wilson, 2010). More specific studies highlight the importance of reducing marketing, transport, and other intermediary costs in agricultural supply chains (Balat, Brambilla, and Porto, 2009; Diop, Brenton, and Asarkaya, 2005). Although

these studies point out the relevance of increased donor assistance to trade facilitation, they do not help delineate the policies and programs that would be most effective in cutting trade costs.¹¹

In their recent authoritative survey of the state-of-the-art literature on industrial policy, Harrison and Rodríguez-Clare (2010) conclude that empirical evidence on the effectiveness of various forms of industrial policy is scarce. The authors look at the case of East Asian countries where industrial policies based on use of production subsidies, subsidized credit, fiscal incentives, and trade protection to foster particular sectors. From this, they claim that the available evidence does *not* answer the most important question: what was the effect of these industrial policies relative to the counterfactual situation where such intervention was absent.¹² In sum, there are no studies that can credibly credit industrial policies with bringing about East Asia's successful industrialization experience. But the authors do make a tentative argument that industrial policies played a role in some countries' growth experiences based on two complementary ideas. First, the composition of a country's export basket — a tilt towards manufacturing or skill-intensive goods rather than primary products or raw materials — seems to matter for its long-run growth. Second, China's export basket in 1992 was much more sophisticated than what would be expected given the country's per capita GDP and that could only be the outcome of its industrial policies (Rodrik, 2006).¹³

Harrison and Rodríguez-Clare's literature survey concludes with an advocacy statement on the type of national trade-related assistance likely to be most successful: that which increases exposure to trade (such as export promotion) in contrast to that which limits trade (such as tariffs or domestic content requirements).¹⁴ The authors also make a statement on the specifics of policy design, where they envision an increasing role for "soft" industrial policies that deal directly with coordination problems, such as those that keep productivity low in existing or emerging sectors. These policies include programs "to help particular clusters by increasing supply of skilled workers, encouraging technology adoption, and improving regulation and infrastructure" (Harrison and Rodríguez-Clare [2010] p. 4112).¹⁵ The problem with this statement is that it has a "rabbit-out-of-the-hat" aspect because the survey includes little supporting evidence. In fact, the absence of evidence for the policy recommendations the survey offers is a reason for our effort to initiate new research on these issues.

¹¹ For example, Portugal-Perez and Wilson (2010) estimate the impact of aggregate indicators of "soft" and "hard" infrastructure on the export performance of 101 developing countries over the 2004–2007 period. Their estimates show that trade facilitation reforms - particularly investment in physical infrastructure and regulatory reform to improve the business environment - improve significantly export performance. Moreover, they show that the marginal effect of infrastructure improvements on exports appears to be decreasing with per capita income.

¹² One empirical approach that has been followed in some studies is to examine whether the sectors that received most support from industrial policies are those that have grown most rapidly; but that approach does not address the counterfactual issue.

¹³ This finding was based on the measure of sophistication of a country's exports basket developed by Hausman, Hwang, and Rodrik (2007) constructed using the level of GDP per capita associated with exports of different goods worldwide.

¹⁴ The authors make this statement based on extensive cross-country and cross-sector evidence on trade and growth.

¹⁵ The authors argue that an advantage of such "soft" industrial policies is that they are generally compatible with the multilateral and bilateral trade agreements that developing countries have entered into in the last decades.

3.3 Evaluating World Bank trade programs

In principle, World Bank trade-related projects should be a key source of evidence on the effects of specific trade interventions, which could become the basis for further evidence-based policy advice. In practice, though, this is rarely the case. Few interventions have undergone rigorous impact evaluation.

An evaluation of World Bank financed trade-related assistance during the 1987–2004 period conducted by the Independent Evaluation Group (IEG) concluded that it helped countries liberalize their trade regimes — average tariffs fell and coverage of nontariff barriers diminished — with positive effects on economic growth (IEG, 2006). However, the evaluation also argued that assistance fell short of generating a strong export supply response. Many client countries, especially in Africa, could not diversify their exports and remained vulnerable to commodity price shocks.

IEG (2006) also discusses the performance ratings of World Bank aid-for-trade projects, which give a sense of their effectiveness in achieving their stated goals. The report shows that trade-related adjustment loans until 2004 performed better than other adjustment loans; whereas trade-related investment loans performed worse than other investment loans of the World Bank.¹⁶ Moreover, according to the same evaluation, assistance on trade logistics — ports, customs, and trade finance — and export incentives had a mixed record, though one that improved over time.

A review of the IEG ratings of recent investment projects and programs on trade promotion, completed as of 2007 (World Bank, 2009), indicates that more than 85 percent were rated as having moderately satisfactory, satisfactory, or highly satisfactory outcomes, which was higher than for projects in other areas.¹⁷ Aid-for-trade projects also had higher estimated economic rates of return (around 32%) than other non-trade related projects (around 23.7%).¹⁸

While providing valuable insights, the IEG evaluations of trade assistance offer limited evidence to support focused trade interventions. Moreover, the evaluation does not cover much of the recent increase in AfT assistance for export promotion and trade facilitation.

¹⁶ Projects that focused primarily on trade liberalization achieved the best performance ratings whereas those related to private financing (such as export finance guarantees and export reinsurance) were the least successful. The superior performance of projects focusing on trade liberalization is not surprising as it reflects the relative legislative ease of putting in place the associated actions (e.g., reform of the tariff regime). In contrast, projects that focused on thematic areas related to key supply-side constraints that impose greater demands on institutional and administrative capacity, such as trade financing, are more difficult to implement.

¹⁷ IEG assesses the performance of roughly one World Bank project out of four (about 70 projects a year) measuring outcomes against the original objectives, sustainability of results, and institutional development impact.

¹⁸ An economic rate of return is the discounted interest rate that would keep an agent indifferent between the choice of undertaking or not undertaking the project.

In search of evidence on the impact of such trade interventions, we conducted a thorough review of the evaluation methods for 85 World Bank trade-related investment lending projects undertaken during the 1995–2005 period. The source of data was the World Bank’s Operations portal website and in particular the Project Appraisal Documents (PADs) and the Implementation Completion Reports (ICRs).¹⁹ The evaluation methods used can be classified into five distinct categories: (a) only economic or financial internal rates of return, net present value or effectiveness calculations; (b) beneficiary surveys and stakeholder workshops; (c) both a and b, (d) both a and b, with a comparison of beneficiaries to a control group; and (e) no formal evaluation methods used.²⁰ One key aspect to note is that the implementation of a beneficiary survey does not guarantee that a rigorous impact evaluation can be conducted since in most cases the survey covers only outcomes pertaining to beneficiaries of the project, and no control group is covered (more details on these methods will be provided in section 4).

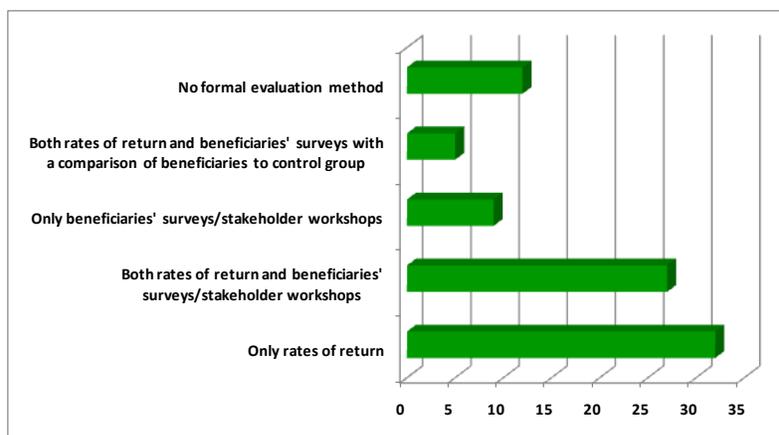
Figure 3 shows that evaluation using only economic or financial rates of return was the most commonly used method for the trade-related projects, while 10 percent of the projects involved no formal evaluation method.²¹ Included in the latter category is a trade competitiveness project that described the impact of the project in purely subjective terms: “While the impact on the firms assisted had not yet been determined, a visit to two beneficiaries by a supervision mission confirmed that there had been an impressive impact on the firms’ quality of products and skills.” Another example of the latter is a trade competitiveness project where the achievement of the overall goal was measured in terms of the higher average annual growth rate of exports during the project duration and increases in exports’ share of GDP compared with the initial year of the project.

Figure 3
Evaluation of World Bank trade-related projects 1995–2005

¹⁹ We thank Vivian Agbhega for compiling the data for this review. The selection of trade-related projects followed the criteria used by Steven Gunawan in an unpublished study of “Monitoring and Evaluation Lessons of Trade Projects” that served as background work for the 2011 World Bank Trade Strategy. The projects were filtered from the World Bank’s Operations portal website according to the theme “Trade and Integration,” and falling within the following criteria: i) approved only after 1995 due to obsolescence; ii) IBRD/IDA-funded; and iii) closed. A total of 321 projects were filtered, out of which 144 were development policy loans and 177 were investment loans, and 30 investment lending projects had to be dropped since they lacked ICRs. A final set of 85 investment lending projects was obtained after excluding projects that did not have any trade components. The main documents used to extract information on the projects were PADs and ICRs. For each project we collected information on the types of intervention, the types of outputs and outcomes achieved, the evaluation methods employed, and the evidence or proof of causation of the impact of the project.

²⁰ A beneficiary survey consists of a formal survey of the entities that received assistance from the project, whereas a stakeholder workshop is a more informal way to collect information on the various entities affected by the project.

²¹ Our analysis of project ICRs did reveal, however, that the use of these methods is often handicapped by difficulty in quantifying some of the costs and benefits of the project. Some project ICRs explicitly say that certain benefits are not incorporated in net present value calculations due to their complexity.



Source: Authors' calculations based on data from the World Bank Operations portal website.

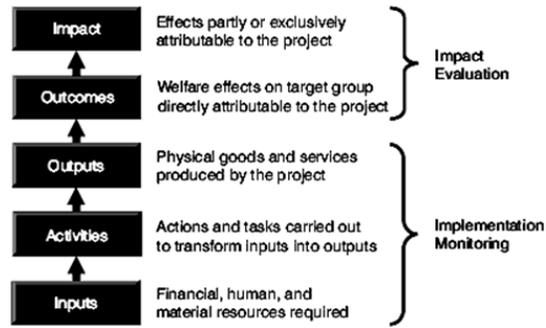
To be fair, task managers of trade-related projects are often candid about their project's achievements, writing in the ICR that observable results (particularly those relating to aggregate outcomes such as total exports) are not entirely the result of the program alone but rather the result of the work and resources of different institutions and sectors. The most striking fact in Figure 3 is that only 6 percent of projects (5 out of 85 projects) included a rigorous impact evaluation, involving a proper comparison of the outcomes of project beneficiaries with those of a control group. But even in such cases, the impact evaluation method raised certain issues, which we will discuss in the next section.

A clarification should be made at this point concerning the link between evaluation methods of projects and the monitoring and evaluation (M&E) framework.²² M&E is an important part of the design and implementation of World Bank lending projects and is the reason why, as mentioned at the beginning of this section, we would expect to obtain evidence on the effects of certain types of trade interventions from project-level analysis. M&E is based on performance indicators capturing outputs, outcomes, and impact of a project (discussed in ICRs).

These categories of performance indicators are thought to be related according to the scheme shown in Figure 4. The scheme makes clear the distinction between considering outputs in general, and going one step further and also considering outcomes and the impact attributable to the project per se. One common concern with the M&E framework for World Bank projects is that it often focuses on the monitoring part and not enough on the evaluation part. For example, most projects include exports as impact indicators but do not include a proper impact evaluation strategy that allows for attribution to the project of an increase in exports.

Figure 4
From inputs to impact

²² This discussion draws heavily on the aforementioned unpublished study by Steven Gunawan.



Source: Adapted from a presentation by Savedoff (2006).

In addition to World Bank investment lending projects, the World Bank also produces a large amount of analytical work — economic and sector work (ESW) — where one could expect to find evidence that supports certain trade interventions. The key trade-related analytical pieces — diagnostic trade integration studies (DTIS) — do highlight the high costs of producing goods and services for export, and for delivering them to foreign markets, as being the major barriers to trade integration in less developed countries, and point to infrastructure as the most pressing constraint. But they do not inform the development community about which interventions work and which do not, and which interventions are most cost-effective.

4. Impact evaluation of trade interventions

The key problem that IE addresses is *attribution* — making sure that observed changes in outcome variables are caused by the program or policy under evaluation and not by outside influences. Many outside influences can confound the identification of a program or policy’s impact. For instance, an export promotion scheme put in place in 2007 would see its positive impact confounded by the negative impact of the global crisis of 2008–2009; a simple before-after comparison of outcomes would likely suggest a *negative* impact of the program.

In order to filter out these influences, one would want to know how beneficiary firms would have performed in the absence of the program (presumably worse). But the data needed for this counterfactual does not exist, because firms cannot be both beneficiaries and non-beneficiaries at the same time. This missing data problem is solved by using as a counterfactual the performance of other firms that did not benefit from the program. By analogy with first agro-biological and then medical sciences, where IE methods originate, beneficiaries are called the *treatment group* and non-beneficiaries the *control group*.²³

²³ A pedagogical reference to IE techniques can be found in Khandker, Koolwal and Samad (2010), which contains analytical guidance as well as case studies and Stata do-files. A formal treatment can be found in Ravallion (2008), Blundell and Costa Dias (2009), and Imbens and Wooldridge (2009).

The central idea of IE is best illustrated by a widely-used technique called double-differences or difference-in-differences. Under that technique, the effect of a program is assessed by comparing the performance of beneficiary firms before and after the treatment (first-difference), and then benchmarking that difference by comparing it to the difference in performance over the same period of non-beneficiary firms (difference-in-differences).²⁴ In our earlier example of an export promotion scheme put in place just before the onset of the crisis, its confounding effect would be captured (and thus filtered out) by the decrease in the performance of non-beneficiary firms during the program period. The program’s impact would then be measured by how much less badly beneficiary firms performed than non-beneficiary ones.

As noted earlier, IE design relies on less-than-universal coverage, which provides a first categorization of programs into targeted and non-targeted ones. Another useful distinction is whether an evaluation is built into program design. In what follows, we consider each of the cases defined in Table 2 in the trade context, and discuss to what extent IE methods can be applied to them. Anticipating our conclusions, our basic argument is that the scope for IE in trade assistance projects is broader than might appear at first, provided that one is not wedded to a particular methodology (randomized control trials for instance).

Table 2
Boundaries of impact evaluation

	Evaluation built into program design	Evaluation not built into program design
Targeted (typically trade competitiveness-related e.g., matching grants for producers for technology upgrading or export business plans; export credit guarantees for producers)	RCT is feasible; Quasi-experimental methods are a possible alternative	RCT is infeasible; Quasi-experimental methods are feasible
Non-targeted (typically trade facilitation-related: e.g., customs reform, port improvements; but also some trade competitiveness -related: support for producer organizations or other institutional reforms)	RCT is typically infeasible; Quasi-experimental methods are more appropriate; Some methods of targeting can be introduced (phase-in, staggered implementation)	All IE methods are difficult; before-after comparisons may be only alternative

Notes: RCT: Randomized control trial; Quasi-experimental methods are matching, difference-in-differences, instrumental variables, or regression discontinuity design.

²⁴ This difference in performance is not a *ceteris paribus* effect: it picks up both direct program effects and induced behavioral changes, which may work to either reinforce or weaken the program’s direct effect. For instance, a program combining matching grants with technical assistance targeted at particular operations within the firm can trigger broader management improvements (a reinforcing influence) or partial waste of program money through management slack (a mitigating influence). See Duflo, Glennerster and Kremer (2008) for a discussion.

4.1 Targeted interventions

Targeted trade interventions include “clinical” trade competitiveness programs such as export promotion schemes through matching grants for supporting export business plans, through export-credit guarantees, or through firm-level technical assistance for technology upgrading, for acquisition of international quality certifications or to meet other product standards. The key feature of these interventions is that the programs are assigned exclusively to certain units, often firms. Because these interventions operate at the level of the firm, non-assisted firms can in principle serve as the control group.

4.1.1 Randomized-control trials

In targeted interventions, *when evaluation is built into program design*, a randomized-control trial (RCT), sometimes called the “gold standard” of IE, tends to be viewed as the best option, though this can be questioned as discussed below. It consists of drawing beneficiaries at random from a large pool of firms. By the law of large numbers, the average characteristics of beneficiaries will be the same as those of non-beneficiaries. Were this condition not met, there would be a selection bias; that is, the program’s impact would be confounded not by outside factors, as before, but by differences in individual characteristics.²⁵ Random assignment to the program ensures that the “unconfoundedness assumption” is verified which is key to identify the average treatment effect (Imbens and Wooldridge, 2009).

Despite its analytical appeal, randomization must confront other difficulties, in general and in the context of trade-related assistance in particular. In terms of practical feasibility, randomization can be a hard sell with client governments for ethical or political reasons. Governments may be reluctant to extend assistance only to a subset of agents when all need it, and any *de facto* discrimination may be politically costly.

Randomization does allow for flexibility, which may help make it acceptable. First, it does not need to cover all individuals. For instance, a program can use standard selection methods to determine eligibility, and introduce randomization among either all eligible firms or only “marginal” ones. That is, very strong candidates can be taken in, very weak ones left out, and only those in the middle subject to randomization.²⁶ Lotteries are somehow more appealing than blind randomization because they avoid the impression that something is hidden. At a more basic level, in the presence of rationed resources to fund a policy intervention, the advantage of randomization is that it constitutes the fairest solution to rationing (Ravallion, 2008).

²⁵ Put differently, the probability of getting treatment, conditional on the individual’s characteristics, needs to be independent of the outcome.

²⁶ We are grateful to David McKenzie for pointing this out to us.

Duflo, Glennersten, and Kremer (2008) note that the spread of RCTs in health, education, and poverty programs owes much to the collaboration with NGOs, as collaboration with local authorities is still relatively rare. NGOs are much less involved in trade-related programs than in other programs, so the scope for RCTs may be inherently less, at least as long as the evaluation culture remains rare in public policy. Atkin and Khandelwal (2011) discuss how carrying out an RCT within the context of international trade depends crucially on finding a suitable local project partner who can provide the export promoting services to producers, and on convincing that project partner of the feasibility and value of the randomization procedure. However it should also be noted that working with NGOs generally limits the size and scope of the intervention and the impact of the program could differ if it is scaled up and implemented without NGO collaboration (Ravallion, 2008).

Atkin and Khandelwal (2011) describe an ongoing project for an RCT to assist microenterprises in the handloom weaving sector in Akhmeem, Upper Egypt to enter into export markets. The project's objective is to link those microenterprises to foreign buyers in the U.S. through the provision of three kinds of services. The first consists of putting Egyptian producers in contact with design consultants to develop patterns that can appeal to the tastes of U.S. consumers; the second is marketing assistance with U.S. buyers; and the third is general business training. The project's impact-evaluation design is simple: after drawing up a list of potentially viable producers/exporters in the sector and region, a random group of them will be given the opportunity to export to the U.S. market with the help of the three services listed above. The data on both outcomes (export performance) and covariates (producer characteristics) will be generated through surveys conducted as part of the IE. A baseline survey will collect information on all viable exporters — both those that will benefit from this intervention as well as those not approached before the services are provided. Another survey will be conducted long enough after the intervention in order for the effects to be tangible.

The World Bank is considering implementing RCTs in some of its own projects as well, although plans at this stage are preliminary. Candidate projects include a customs border post modernization project at the border between the Democratic Republic of Congo and Rwanda, where petty traders on foot, mostly women, are regularly exposed to corruption and harassment. The project would involve some of the women's associations (with group randomization) to designate customs brokers acting as shields between women and predatory customs officers. Another project involves the facilitation of payments for small cross-border transactions through branchless banking near the Cameroon-Chad border. Currently, all payments for such transactions are made in cash, which hampers trade. At the very least, the project would involve a natural experiment if branchless banking is allowed for traders on one side of the border but not on the other; in addition, the design may involve, in a pilot phase, selected access to non-cash payments for a randomly chosen treatment group.

One of the reasons why RCT is a preferred design for such experiments is that randomization does away with the need for complex econometric techniques to control for selection in non-experimental settings. However, RCT is no silver bullet in small sample environments, as it relies on the law of large numbers to ensure that expected untreated outcomes are equal in treatment and control groups. In low-income countries, interventions sometimes target very small numbers of firms (McKenzie, 2011b).²⁷ For instance, the Pesticides Initiative Program (PIP), an E.U. technical-assistance program designed to help fruit and vegetable producers cope with E.U. standards, covers less than a few dozen firms in some African countries (Jaud and Cadot, 2011). Randomization is not an option in such environments. Quasi-experimental methods may not do very well either; but if a cross-country sample is available with enough observations, econometrics may offer some scope to control for cross-country heterogeneity.²⁸ We will return to the small sample issue in the context of non-targeted interventions in section 4.2 when referring to the Cameroon customs project described in Cantens, Raballand, Bilangna, and Djeuwo (2011).

An intrinsic limitation of RCTs in the trade and other economic areas is that the study subjects are active economic agents who consciously choose their responses as opposed to the medical sciences where passive entities (e.g., cancer cells) respond endogenously following the laws of nature (Barrett and Carter, 2010). Unobservable perceptions about the benefits of a new trade-related intervention will vary among potential beneficiaries in ways that are likely to be correlated with other attributes and with the actual effects of the treatment. In section 6, we will discuss how randomization may fail to produce unbiased treatment effects in the presence of “essential” heterogeneity or in the presence of spillovers.

4.1.2 Quasi-experimental methods

When evaluation is not built into program design, RCT is not an option and quasi-experimental (QEM) methods must be used, all relying on econometric techniques to overcome selection bias.²⁹ The first is the difference-in-differences (DID) method briefly described above. By comparing differences in outcomes instead of comparing levels, DID controls for unequal performance levels of treatment and control groups not related to the program. However, DID

²⁷ McKenzie (2011b) discusses the issue of small samples in World Bank private sector support programs in Africa. None of those programs has been subject to rigorous impact evaluations so far, but if such evaluations were to be conducted researchers would be faced with a serious problem of power given the small number of enterprises assisted by the projects and their large degree of heterogeneity.

²⁸ Randomization across countries would be more difficult to implement than within a country and would not necessarily increase the test’s power.

²⁹ How well quasi-experimental methods perform compared to randomization has been a subject of intense scrutiny since the seminal paper of Lalonde (1986), with largely inconclusive results. Glazerman, Levy and Myers (2003) found that quasi-experimental methods produced substantially biased results compared to experimental ones in twelve replication studies of welfare and employment programs in the U.S. Cook, Shadish, and Wong (2006) found less clear-cut results for education programs. See Ravallion (2008) on the evaluation of poverty programs in non-experimental settings.

relies on the assumption of parallel trends and does not control for selection on observables (firm-level covariates).

The DID method can be improved by matching that controls for observed firm characteristics correlated with both program participation and performance. The key assumption for the impact estimated by this method to be unbiased is that selection into the program is based only on observable firm characteristics.³⁰ The matching procedure evolves in two steps. First, firm-level covariates are used to predict the probability of getting (or enrolling into) the program using a probit or logit regression. This predicted probability is called a propensity score. Second, the control group is formed by picking, for each treated firm, the untreated firms with the closest propensity score. For each treated firm, depending on the method, there can be either one matched control firm or several, using a weighted scheme.³¹ Average outcomes in first differences are then compared between the treatment group and the matched control group. The propensity score matching DID estimator allows for time-invariant unobserved firm heterogeneity to affect selection and outcomes. But it does not address the problem that selection - as well as outcomes - may depend on unobserved time-varying firm heterogeneity, as will be discussed below as well as in section 6.3.

The studies surveyed in Volpe (2011) are good illustrations of the use of quasi-experimental methods in the evaluation of trade assistance. These studies, recently carried out at the Integration and Trade Sector of the Inter-American Development Bank, use DID and matching-DID methods to assess the effectiveness of export promotion activities of PROMPEX/PROMPERU (Peru), PROCOMER (Costa Rica), URUGUAY XXI (Uruguay), PROCHILE (Chile), EXPORTAR (Argentina), and PROEXPORT (Colombia). They use rich and unique datasets for the six Latin American countries that combine firm-level customs data with covariates drawn from other national firm-level data sources, and constitute the first rigorous micro-based evidence of the effects of export promotion.³² The picture emerging from Christian Volpe's survey is that export promotion was effective in facilitating export expansion for firms in the LAC region, but primarily along the extensive margin. Firms exporting differentiated goods benefit more than those selling more homogeneous goods. Small and relatively inexperienced companies benefit more than larger and already established exporters. Finally, bundled services that provide support to firms throughout the export development process appear to be more effective than isolated actions.

³⁰ This assumption is designated as "ignorable treatment assignment" by Rosenbaum and Rubin (1983) which is the seminal study on propensity score matching estimation. The assumption means that program participation and outcomes are independent, conditional on a set of observed attributes.

³¹ The single-match method is called "nearest-neighbor." Alternatively, one can use n nearest neighbors, or the entire sample of untreated firms with weights that decrease with distance from the treated firm's propensity score. This latter method is called "kernel matching." Many other refinements are possible. See Caliendo and Kopeinig (2005) for details on propensity score matching estimators.

³² An alternative, more traditional route to the evaluation of export-promotion's effectiveness is the aforementioned cross-country study of Lederman et al. (2010).

Gourdon, Marchat, Sharma, and Vishwanath (2011) apply the same type of quasi-experimental methods to the evaluation of FAMEX, a World Bank-supported export promotion program in Tunisia, which provided a mixture of counseling and matching grants to new exporters. The study exploits a customized firm-level survey to estimate the effects of FAMEX on the export performance of beneficiary firms at the intensive and extensive margins. Propensity-score matching DID estimates suggest a very large and statistically significant growth effect at the intensive margin: a 39% differential in terms of annual export growth compared to control firms over the 2004-2008 period. The treatment effect at the extensive margin – in terms of products and destinations – is both smaller quantitatively (a 5% growth differential in the count of products and destinations for program beneficiaries compared to control firms) and of marginal or no significance (at 10% confidence level for destinations and insignificant for products). In addition to the observed acceleration in export growth, Gourdon et al. find a significant boost to employment growth: a 10% annual differential for program beneficiaries, significant at the 5% confidence level. An original feature of their dataset is that it covers service firms in addition to manufacturing firms, and they find considerably stronger effects for the former. One potential issue with their data is that the survey was conducted ex-post (no baseline survey was conducted as IE was not part of the program design) so the data may suffer from recall bias. Preliminary results in Cadot, Fernandes, Gourdon, and Mattoo (2011) based on an alternative source of data (customs data) suggest a smaller and non-persistent treatment effect.

Jaud and Cadot (2011) also apply quasi-experimental methods to assess the impact of the E.U.-funded pesticides initiative program (PIP) on the export performance of firms in Senegal's horticulture sector. Their results suggest that, while the program had no significant effect on exports of fresh fruit and vegetable pooled over all products and destinations, it had a positive effect when considering exports to the EU.

Other quasi-experimental methods can address selection bias in the evaluation of the impact of a program. One approach relies on instrumental variable (IV) estimation. This can be used when program take-up is less than complete and thought to be correlated with unobserved individual characteristics influencing performance. In this case, eligibility can be used as an instrument for participation, provided that eligibility is truly exogenous (e.g., if there is randomization of eligibility but program take-up is incomplete or some participants drop out). This method is used in the context of non-targeted interventions by Sequeira (2011), as described in section 4.2.

Another approach is regression discontinuity design (RDD), which makes use of breaks in eligibility to identify a program's impact.³³ For instance, suppose that an export promotion program targets small and medium-sized enterprises (SMEs) as defined by a cutoff level of sales. If the sample is large enough, one can compare outcomes for SMEs immediately below the

³³ See Campbell (1969) for details and a survey can be found in Todd (2008).

cutoff (eligible) and for SMEs immediately above (ineligible), on the assumption that they are close enough in the characteristic upon which eligibility is defined to be good matches for each other, and most importantly that the cutoff rule is indeed enforced.³⁴

4.2 Non-targeted interventions

Non-targeted trade interventions cover mostly programs that help reduce trade costs. These include trade facilitation programs such as upgrading of bottleneck infrastructures in ports, roads, or railroads, reforms of customs agencies and procedures, and some types of trade competitiveness programs related to general improvements in the business environment or support to producer organizations. Because these interventions generally do not target micro entities and their direct beneficiaries are multiple and diffuse, the identification of a control group is difficult, and so they are less amenable to experimental or quasi-experimental design.

Considering "hard" and "soft" infrastructure-related trade facilitation programs, the two key constraints to estimating their effects are (a) the endogeneity of program placement and (b) the absence of well-defined treatment and control groups. Thus, the pre-treatment unobservable characteristics that determine infrastructure placement and affect outcomes will likely differ between treatment and comparison groups (where groups are, in this case, most likely to be locations). Randomization in the context of large and sensitive hard transport infrastructure programs is generally not feasible. This is also the case for soft trade facilitation programs relating to rules, regulations, and government agencies dealing with the movement of cargo across borders that are often not amenable to random assignment at the micro-level nor to the creation of comparison groups for the purposes of an IE.

For interventions such as customs reform, the only way to generate a control group is to introduce elements of targeting through progressive phase-in during a pilot phase, staggered for example across different border posts, or through selective implementation covering only some customs offices or officials, or by giving privileged access only to some firms or to some types of traded goods. For instance, a "green channel" in customs, which is a speedy clearance for trusted operators, can be restricted and randomly allocated in an early phase, using non-eligible operators as controls.³⁵ In this case, methods such as DID can in principle be applied using the locations initially not covered, customs offices or officials, or firms like the control group for the targeted entities.

However, in many cases, during the pilot phase the control group will not be strictly comparable to the treatment group. For example, when a border modernization program is initially deployed

³⁴ The issue of rule enforcement has been a controversial one for example in the context of microcredit evaluation (see Morduch 1998), but may be lesser concern for firm-level trade interventions such as support to SMEs.

³⁵ This approach is similar to so-called "pipeline" methods where applicants are used as controls for beneficiaries.

in one border post, other border posts of different scale and product mixes serving other areas could serve as controls. It may then be necessary to use regression analysis to control explicitly for the heterogeneity in covariates in estimating differences in outcomes between treated and control border posts.

In some cases, policy design or implementation inadvertently creates the conditions necessary to perform evaluation through quasi-experimental methods — what economists call a “natural experiment.” Datt and Yang (2011) exploit one such natural experiment. The government of the Philippines used pre-shipment inspection (PSI) services to combat corruption in customs and increase import duty collections. The natural experiment arose from two conditions: (1) imports from only *some* origin countries were covered by PSI, which created a natural control group (imports from other countries); and (2) in 1990 the government decided to close a loophole whereby import transactions below a threshold of \$5,000 were exempted from PSI. The loophole had enabled traders to slice shipments into small batches and under-invoice them without being detected. The customs reform consisted of lowering the threshold to \$500, so the period after 1990 can be considered a “treatment period.” A DID equation can then be used to compare the evolution of outcomes before versus after the reform for the treatment and control group of countries. The DID estimates show that, when inspections were expanded to lower-valued shipments, imports shipments were no longer mis-valued, but those from treatment countries shifted differentially to an alternative duty-avoidance method — shipping via duty-exempt export processing zones (EPZs). Thus, increased enforcement reduced the targeted method of duty avoidance, but led to substantial displacement to an alternative duty-avoidance method. Duty collection failed to rise, while importers incurred higher fixed costs as they relocated to EPZs. This evidence shows that, to be successful, anti-corruption reforms need to encompass a wide range of possible alternative methods of committing illegal activity.

Sequeira (2011) discusses a transport infrastructure project consisting of investments in a railroad connecting the economic heartland of South Africa to the port of Maputo in Mozambique. Given the poor state of Mozambique's infrastructure after two decades of war, and in face of budget constraints, the government had to be selective in its choice of infrastructure investments. They decided to rehabilitate the old-pre-colonial railway in the Maputo transport corridor (that would promote regional integration) rather than building an entirely new North-South connection as was demanded by the Mozambican business class. As the layout of the old-pre-colonial railways had been designed to serve 19th century mining companies, there is plausible exogenous variation in the emergence of the rehabilitated railway relative to the geography of manufacturing and retail firms at the time of rehabilitation of the railway.

The IE of this transport infrastructure project estimates the impact of railway rehabilitation on firm performance — namely, how it affects transport costs for different firms and sectors, how firms respond to these changes, and what are the spillover and network effects across rail and

road transport. To identify a causal relationship, the study will use a quasi-experimental method, IV, where the treatment, defined as changes in transportation costs, will be instrumented by the distance between a firm's location and a working station of the railroad.

In addition, the study exploits the fact that other transport corridors in Mozambique developed at different speeds and identifies two sets of control firms to match to the treated firms in the Maputo transport corridor: firms in the Beira corridor (that have access to a new port but no railroad) and firms in the Nacala corridor (that have no access to a new port or railroad). To isolate the impact of the Maputo railway rehabilitation, the study will use a matching DID estimation that assumes that the only factor making the trajectory of these three sets of firms different during the sample period is that they were exposed to different transport choice sets. The impact of the Maputo railway rehabilitation is not yet known since only the baseline survey information is available; a follow-up survey will be conducted in 2011.

Sequeira (2011) also discusses a "soft" transport infrastructure project focusing on corruption in Southern African ports.³⁶ By collecting original data on bribe payments made to customs officials and to port operators in the two competing ports of Durban and Maputo, the study is able to trace differences in bribe schedules to the organizational structure of each port. By observing how firms adapt their shipping and sourcing decisions to the type of corruption faced at each port — which enters the calculation of the overall cost of using each port — the study estimates the impact of corruption at ports on the behavior of South African firms. The estimates show that corruption imposes a distortion in terms of “diversion” — i.e., firms travel on average an additional 322 kilometers, more than doubling their transport costs, just to avoid “coercive” corruption at a port.

This effect is only observed for firms facing a higher probability of being coerced into a bribe because of the kind of product they ship. Firms are willing to incur higher costs to avoid corruption because of an aversion to the uncertainty surrounding bribe payments at the most corrupt port (Maputo). The uncertainty in Maputo seems linked to the short time periods caused by high job turnover among customs officials. Firms also respond to different types of corruption by adjusting their sourcing decisions for inputs — domestically or internationally — since corruption at ports increases the cost of using the port and thus directly affects the relative cost of imports.

While this project is not an impact evaluation of an intervention to reduce corruption in ports, it provides two sets of valuable insights on such interventions because it considers the entire chain between competing port bureaucracies setting bribes and user firms making shipping and sourcing decisions. First, the study shows that, depending on the type of corruption that

³⁶ The project is described more extensively in Djankov and Sequeira (2010).

bureaucrats engage in, bribes can affect the deadweight loss, tariff revenue, and the demand for the public service. In particular, corruption seems to reduce significantly demand for the Maputo port, stifling the returns to the massive investments in hard infrastructure of the corridor that have taken place in recent years. Second, policy changes to the organization of ports and to the nature of the interaction between shippers and port officials could reduce corruption. Such changes include reducing the discretion of port officials in the clearance process, and eliminating face-to-face interactions between clearing agents and port officials.

Cantens, Raballand, Bilangna, and Djeuwo (2011) describe a recent pilot for customs reform in Cameroon that involved the introduction of contracts with performance indicators for frontline customs inspectors in two of the country's customs bureaus (henceforth referred to as treated bureaus). The performance indicators covered both trade facilitation and the fight against fraud and bad practices. Frontline customs inspectors with good performance would be rewarded with non-financial incentives such as congratulatory letters entered into their personnel files, easier access to the director general of customs, training courses, and transfers to more attractive bureaus. Poorly performing inspectors would be sanctioned by eviction from bureaus with strong "fiscal potential" — that is, where the possibilities of earning money legally through disputed claims were high.

This project is an interesting example of a trade intervention that in principle is non-targeted, but where targeting could have been introduced by focusing on a sub-set of frontline customs inspectors. This could then have been an ideal setting to implement an RCT, whereby a subset of randomly chosen frontline inspectors would have been under performance contracts while others would not. However, it was not possible to implement an RCT for several reasons. First, the seven customs bureaus in Cameroon are specialized (oil imports, special customs regimes related to public trends, transit, exports, bulk cargo, and the two treated bureaus) and differ so much in customs practices that it would be difficult to make comparisons across bureaus. Hence, if anything, one would need to take a bureau and split it into a treated group and a control group of frontline inspectors. But this was not feasible given a small sample problem: less than 10 staff work in each bureau. Second, as is generally the case in projects funded by governments or international donors, the time for the pilot project was limited. Thus, it was not possible to overcome the small sample issue by allowing for turnover within each bureau to artificially increase the number of treated and control officers. Moreover, since contract incentives were not financial, time was required to reward good performers (e.g., it would not have been feasible to appoint high-performing inspectors to better positions every six months).

Therefore, the IE of the customs performance contracts project was conducted as a comparison of inspectors' behavior before and after the project was implemented, without a defined control group, although the impact on clearance times was assessed using the bulk-cargo import bureau

as a counterfactual. The estimated effects of the pilot performance contracts were positive surprisingly soon after the pilot was launched in mid 2009. Duties and taxes assessed increased despite a fall in the number of imported containers (likely linked to the financial crisis), and the tax yield of the declarations also rose. The performance contracts also affected clearance times, as the share of declarations treated within 24 hours increased more in the treated bureaus than in the counterfactual bureau, and the variance of clearance times decreased dramatically. The impact on disputed claims was equally interesting, with inspectors abandoning low-level disputed claims to focus on major ones, and the ratio of taxes adjusted to taxes assessed increased. Finally, the contracts also had a major impact in reducing costly practices. For instance, the number of litigious re-routings from the yellow channel (documents control) to the red channel (physical inspection) declined tremendously.

5. Data issues

In this section, we discuss first, how the objectives of the evaluation influence the type of performance measures that need to be considered, and then how the necessary data may be obtained.

5.1 What should we measure?

The choice of performance measures is important not only to ensure that IE focuses on the appropriate indicators, but also because using IE can affect the incentives of agents and program managers in unintended ways. Performance indicators that strongly relate to targeted interventions in a causal sense are often too technical to be of interest from a broad policy perspective; whereas, the highly aggregate indicators that interest policy-makers are rarely faithful reflections of the effect of targeted interventions and projects. Thus, selecting performance indicators involves a trade-off between breadth and identification.

Much of the talk in aid-for-trade evaluation focuses on aggregate indicators such as national export performance or other macro variables. Although policy-makers may find these broad indicators relevant, the causal link between them and the actual performance of trade interventions is tenuous, implying weak identification.

By contrast, M&E frameworks, developed to ensure project management and quality control, have used intermediate outcomes more directly linked to the projects themselves, like customs clearance times. In a causal sense, these measures are closer to project management but are likely to be narrow in scope. Deciding which approach is better depends on what the indicators are used for. If evaluation results are expected to feed into incentive structures for program managers, identification is critical and breadth is secondary. In contrast, in order to catch the attention of policy-makers, breadth matters more, possibly at the cost of weaker identification.

Impact evaluation does not escape this general trade-off between breadth and identification, but typically locates at the “narrow” end of the spectrum since it identifies changes in performance measures that are directly attributable to the project. For instance, when evaluating a customs modernization program, the performance measure is likely to be something like container dwell time, even though less quantifiable dimensions of customs performance, like security at the borders, may also matter.

But identifying and documenting the chain of causality from program to ultimate outcomes can be challenging for some trade interventions. In trade facilitation programs, it is not always clear what are the micro-level mechanisms by which transport costs reductions influence firms and households and, more generally, economic activity.

In addition, the use of IE can affect incentives in the long run. The focus on narrow, immediate performance outcomes may well lead to measurement biases or, even worse, create perverse incentives when used for monitoring and evaluation. For one thing, it can focus attention on readily measured outcomes at the expense of less easily measured ones. Consider a customs modernization program. Using IE results to design reward schemes for customs officials might lead to over-emphasis on easy-to-measure reductions in clearance times, at the expense of the monitoring of suspect shipments. If, say, there is a low rate of smuggling illicit products, it may take time before the consequences of reduced monitoring get noticed — too long to show up in an IE.

5.2 How do we obtain the data?

The feasibility of rigorous impact evaluation hinges critically on data availability. Whether the IE is based on experimental (RCTs) or quasi-experimental design, it needs to include a baseline survey and at least one follow-up survey. If quasi-experimental methods are used, the baseline survey must include a rich set of covariates to estimate a (first-stage) selection regression. One of the advantages of RCTs, especially in developing countries, is that they are less demanding in terms of data; however, even with randomization, firm-level covariates can be useful in verifying that the treatment and control groups are comparable in their observable characteristics. This is especially important for small samples. The availability of a rich set of covariates allows for the analysis of heterogeneity in the effects of the program. Moreover, a deep knowledge of the objectives of the program as well as its administrative and institutional details can be important for the design of surveys that collect the right type of information to control for the selection process (Ravallion, 2008).³⁷

³⁷ Qualitative information, collected from surveys or focus groups, can complement quantitative survey data though it cannot be the basis for credible impact evaluation by itself.

Table 3 provides examples of intermediate and ultimate outcomes in the context of new-style trade interventions linked to trade competitiveness and trade facilitation.

Table 3
Intermediate and ultimate performance outcomes

	Trade Competitiveness	Trade Facilitation
	Example of program: matching grant to support firms access export markets	Example of Program: Customs reform
Intermediate outcomes to understand the chain of causality from program to outcomes	Exports, output, input choices at firm-level	Customs or port clearance time and costs, incidence of illegal activity
Ultimate outcomes	Productivity, wages, employment at firm-level	Trade volumes, customs revenue collected
Covariates to use as controls or to understand the heterogeneity of effects of program	Firm-related industry, location, age, size, ownership, workforce details	Firm-related or customs office or official-related: location, education, age, contract

The evaluations of the impact of trade facilitation programs — especially those related to infrastructure — suffer currently from a serious lack of micro-data on transport costs and prices before and after interventions take place. For these types of interventions, it is desirable to conduct baseline and follow-up surveys of program beneficiaries and control groups.

In addition, baseline and follow-up surveys may not be enough to assess a program’s impact. Consider, for example, the case of a one-year export-promotion program, where firms can enlist in any year between 2005 and 2009; and then suppose that a baseline survey is conducted in 2004 and a follow-up survey is conducted in 2010. For firms that enrolled in 2005, the follow-up survey will pick up outcomes four years after the treatment. By then, if the effects are transient, they may have vanished, and the follow-up survey will pick up heterogeneous effects (one year after treatment for firms enrolled in 2009, two years for those enrolled in 2008, and so on). Thus, although costly, it may be necessary to run repeated follow-up surveys year after year.³⁸

While projects typically have budgets for baseline data collection, these may not always be enough to gather the data needed for a proper IE evaluation after the project is completed. An alternative cost-efficient method is to use official pre-existing sources of data provided that they are collected often enough and provided that they can be closely reconciled with program data. For example, in the IE of an export promotion program, customs records at the transaction/firm level can be used to measure outcomes such as growth in export value (the intensive margin),

³⁸ McKenzie (2011a) argues that two advantages of having multiple data points for treatment and control groups are (1) the possibility of studying the trajectory of program impacts and uncovering causal chains and (2) the collection of multiple measurements on possibly noisy and weakly auto-correlated outcomes. By averaging outcomes across multiple data points noise is eliminated and the power to detect genuine effects of a program increases.

number of products, or number of destinations (the extensive margin).³⁹ Naturally it is important to integrate such data with program data such as from the project monitoring database.

The trade and integration unit of the World Bank Development Research Group is involved in a major data collection exercise that may help the IE of trade-related interventions in the next few years. As described in Freund and Pierola (2011), the exercise consists of the collection and compilation of the first ever database on exporter-level customs transaction data across countries and over time. Data has been obtained for 20 countries in Africa, Asia, Eastern Europe, and Latin America, and negotiations are in progress to obtain data for 25 more countries. The database will include statistics on exporters' characteristics and behavior by country, industry, and destination market. The purpose of the database is to provide policymakers, development agencies, researchers, and the public with a novel source of information to conduct analysis of export growth at the micro level and allow for the evaluation of programs and policies affecting that growth.

Data on firm characteristics (covariates), used to control for selection bias, is typically hard to obtain. If an industrial survey is available in the country where the trade intervention is taking place, it can provide the required variables (e.g., location, age of the firm, education of its head, number of employees, foreign ownership). However this requires that customs and industrial census data be merged, which raises confidentiality concerns and may require active collaboration by busy officials in local institutions. When data is not available, an alternative is to conduct a "retrospective survey" — although this method may be biased. In its evaluation of Tunisia's export-promotion agency, Gourdon et al. (2011) use a combination of data from a survey and from national sources (the customs agency and national statistics institute). Yet another alternative is to include questions on program participation and details in ongoing surveys. This also requires close collaboration between the evaluator and the local institution implementing the survey.

6. Looking ahead: Challenges facing IE of trade assistance

In this section, we consider three key challenges that credible IE of trade interventions must address.

6.1 External validity and cost

One concern with impact evaluations is that their external validity is an act of faith. When a program is found to be effective (or ineffective), how do we know that the result would carry over to similar programs run in different environments?

³⁹ See Freund and Pierola (2011) and Lederman, Rodriguez-Clare, and Xu (2011) for uses of such data in a non-IE context.

As Rodrik (2008) and Ravallion (2008) have argued, there is a trade-off in policy evaluation between external and internal validity. As traditional identification of causal effects through instrumental-variable strategies never completely eliminates confounding influences, these strategies always suffer from an internal-validity problem. However, when based on cross-country evidence, they pick up average effects that can be relatively stable — provided they are consistent with some sort of theory — because induction, even on cross-country samples, may fail to produce generalized results.

By contrast, IE purges out confounding influences, but generates results that are empirical and case-dependent. Such results may fail to carry over to different settings. Limited external validity of any study would not be a problem if we could replicate it easily. With enough replications, the sheer mass of evidence would provide the desired generality (although the method would still be inductive and would thus suffer from the general critique of inductive methods in science). But some kinds of IE can be costly. For instance, the World Bank reckons that household surveys cost on average \$300 per household. At that rate, a baseline and final survey of 500 households would cost \$300,000. This is a lot for studies with only internal validity.⁴⁰ However, costs can often be contained by working with local institutions, which has the added advantage of building capacity in a key area.

Some trade-related programs target limited numbers of firms, so their evaluation is less costly than that of poverty-reduction programs. For instance, in a middle-income country, the cost of surveying 500 firms can be substantially lower than \$100,000. Moreover, the data may exist prior to and independently of the IE in the form of census or industrial surveys and customs records. In that case, the cost of the IE goes down dramatically. The problem then is no longer one of cost but more of securing buy-in from the agencies possessing the data so that they share it. However, it should also be kept in mind that tests of the effect of interventions based on 500 firms are likely to have low power (see McKenzie 2011a and 2011b for a discussion), and thus to generate type-II errors (failing to reject the null hypothesis of no treatment effect when, in fact, the effect is present). If cost-cutting leads, through low-power experiments, to unjustified pessimism on the effect of interventions, IE may lose a lot of its power to guide policy choices.

6.2 Spillovers and general equilibrium effects

⁴⁰ In their discussion of quasi-experimental versus experimental methods, Duflo et al. (2008) make a noteworthy point about the commitment value of costly experimental design. It has often been argued, with some statistical support (see, e.g., Ashenfelter, Harmon, and Oosterbeek 1999), that statistically significant results (positive impact in our setting) are more likely to get published, a so-called “publication bias.” As experimental methods are costly and usually planned with donors, self-censure in the face of insignificant results is less likely to be feasible than when relatively low-cost quasi-experimental methods are used with publicly available data. In that sense, IEs may be less affected by publication bias.

Externalities can bias treatment effects by blurring or magnifying the difference in outcomes between treatment and control groups. In the context of policy evaluation, this raises a deep issue, as externalities are often the basic justification for government intervention.

One key assumption of both experimental and quasi-experimental methods is that the impact of the program can be located only among its direct participants, that is, the control group is not “polluted” by the treatment group, lest the comparison of outcomes be biased. A classic case in economics occurs when general equilibrium effects transmit the benefits conferred on beneficiaries to non-beneficiaries or, alternatively, penalize them, say, through rising input prices. For example, a program to upgrade one border post may induce traffic shifting from other, untreated border posts. The volume of trade going through the treated border post will then be increased by the substitution from traffic that normally goes through other posts, and as “control” border posts see their traffic go down, using them as controls will result in an upward bias in the estimated treatment effect of the program. Similarly, beneficiaries of an export promotion program may be able to lure away the ablest workers from other non-beneficiary firms. The adverse effects on the latter would lead to an overestimate of the benefits of the program for the former. Thus, ignoring general equilibrium effects can produce misleading evaluations of policies and programs (Abbring and Heckman, 2007).⁴¹

In the evaluation of trade-related programs of limited scale, such as export promotion or trade facilitation, general equilibrium effects through market mechanisms may not be critical. However, spillovers may be present through other channels – such as social interactions, which are direct externalities in which the actions of one agent directly affect the actions (preferences, constraints, technology) of other agents (Abbring and Heckman, 2007). For example, an export promotion program may have “demonstration effects” yielding valuable information on the viability of products or destination markets that can be easily imitated by non-participants. In such circumstances, the estimated treatment effect will be biased downward because the difference in outcomes between the treatment and control groups will measure only the purely private effect. It is hardly surprising that treatment effects may be contaminated by information externalities. After all, even the most rigorous RCTs used to test the effectiveness of drugs can be affected by informational biases, as is the case when individuals in the control group observe that they do not suffer a drug’s side-effects while individuals in the treatment group do and as a result infer that they received the placebo instead of the treatment. Interestingly, the problems originating in the confounding spillover effects from a program to the control group are as relevant for targeted interventions as they are for non-targeted interventions. In fact, as pointed out by Ravallion (2008), they may be a more severe problem for randomized evaluations.

⁴¹ But Abbring and Heckman (2007) acknowledge that it is costly to obtain information on all the behavioral parameters required to conduct general equilibrium evaluation.

In the trade context, as in economics more generally, the presence of externalities takes on special importance because it plays a key role in justifying government intervention. If the benefits of a program whose costs are borne by taxpayers were internalized by the beneficiaries, surely those beneficiaries ought to pay for it and there would be no justification for public intervention. In contrast, if a program generated spillovers so powerful that no treatment effect was detectable – that is, the control group indirectly benefited as much from the program as the beneficiaries - then there would be a strong argument for a public intervention, as beneficiaries would be willing to pay *nothing* for the treatment.⁴²

These arguments suggest that impact evaluation results cannot be properly interpreted without a careful discussion of what market failure(s) the policies or programs are trying to remedy. Understanding clearly the policy objectives, the relevant constraints (including those related to resources, information, incentives, and political economy) and the causal links through which the specific policies and programs yield expected outcomes are key for any good evaluation (Ravallion, 2009).

If the market failure lies in imperfect capital markets, then a program that provides cheaper-than-market trade financing to specific firms can be expected to have a positive treatment effect and evidence of such an effect can lead one to conclude that the program works. In this case, if there are spillovers, then the positive treatment effect is simply a lower bound of the total effect of the program. However, if the market failure is due to informational spillovers, so that private firms wait for other private firms to invest in uncovering the information needed to export a particular product or to a particular market, then the absence of a treatment effect from an export promotion program is not evidence that the program is not working. In fact, finding a positive treatment effect could reflect the fact that the benefits are largely private, in which case the rationale for the program is put into question.

Thus, seeking to justify government-financed programs solely on the basis of treatment effects may not only be affected by bias, it may be altogether wrongheaded. In the export promotion program example, what IE would be measuring is only the private-good dimension of the intervention; the public-good dimension would be left unevaluated.

It is thus important to disentangle whether a no-effect finding is due to externalities or to program ineffectiveness. This may call for an independent effort, aside from the IE itself, to detect the presence of externalities. For instance, one might estimate a regression of outcomes of untreated individuals on some continuous measure of exposure/closeness to treated individuals, to see if more or closer treated neighbors raise the outcome of untreated ones. Alternatively, one can include this same measure of exposure to (other) treated individuals in the DID equation and

⁴² This point was made to the authors by Daniel Lederman.

interact it with the treatment to see if the treatment is more powerful on individuals “surrounded” (in some economic sense) by other treated individuals. These methods are inspired by measures of contagion used in epidemiologic studies.

In contrast with medical sciences, however, in social sciences the mechanisms by which contagion takes place are largely unknown. Baseline surveys as well as qualitative information gathered from focus group discussions may help in understanding and identifying channels through which future program benefits might spread from one firm to another — for example, professional association memberships, personal contacts and so on.

6.3 Heterogeneity

Differences among beneficiaries, especially if they are unobserved, can pose particular challenges for evaluation. Policy interventions can have diverse impacts across economic agents. By focusing on average treatment effects, an evaluation ignores valuable information on the heterogeneity of the effects.⁴³ As Ravallion (2009, p. 37) puts it, practitioners should “never be happy with an evaluation that assumes common (homogeneous) impact”. He also argues that knowing more about the heterogeneity of the effects and the role of contextual factors is key to better understand the impact of the intervention and make evaluation more relevant for good policy-making. The challenges linked to the heterogeneity of the effects are relevant across types of interventions, whether they are of a targeted or a non-targeted nature.

First, the treatment effects of a program can be related to the observable characteristics of the beneficiaries. For example an export promotion program can have differential effects for participant firms depending on their prior export experience or on their workforce skill levels. If the export promotion program consists of a matching grant scheme which co-finances firms’ export business plan, the opportunity costs for participant firms may differ in terms of the alternative uses they could give to their funds.⁴⁴ A simple approach to address the heterogeneity of the effects when differences are observable is to add interaction effects with the treatment dummy variable in a regression framework that estimates the average treatment on the treated effect. Treatment effects can also vary with the distribution of outcomes themselves. Volpe Martincus and Carballo (2010) examine the impacts of export promotion activities across quantiles of the distribution of Chilean firms’ growth rates of exports using quantile treatment effects estimation. They find stronger effects at the lower end of the distribution, which are combined with data on firms’ export histories to show that smaller and relatively inexperienced firms as measured by their total exports benefit more from export promotion.

⁴³ See Abbring and Heckman (2007) and Ravallion (2011) on the heterogeneity of impacts in evaluation studies.

⁴⁴ While these opportunity costs may not be observable, they are likely to vary with observable characteristics such as the conditions in the local market or in the sector of activity.

Second, a particular challenge arises when unobserved differences among beneficiaries influence their participation in a program. Even in RCTs where eligibility is randomized, economic agents inclined to take up the program based on their unobserved expected net benefits may differ systematically from the agents that were part of the sample randomly assigned to the treatment group. Heckman, Urzua and Vytlačil (2006) introduce the notion of “essential” heterogeneity as pertaining to the case where the impact of a policy intervention is heterogeneous and agents take up treatment based on this heterogeneity (i.e., with knowledge of their idiosyncratic response). The presence of “essential” heterogeneity implies that the estimated average effects of an intervention, even under an RCT, can be biased. For instance, unobservable characteristics of firms that determine their choice of applying to an export promotion matching grant scheme (as well as their choice of the type of business plan and amount of co-finance to provide) could influence the firms’ export success and thus the true effect of the program. The econometric approaches to address “essential” heterogeneity problems are at the frontier of the impact evaluation research (see Manski, 1997; Abbring and Heckman, 2007; Djebbari and Smith, 2008; Fan and Park, 2010).

There are at least two approaches to mitigate the problem of selective take-up on the basis of unobservable attributes. Consider again the export promotion scheme example. First, all firms could be invited to participate in the program, some firms would officially apply, but then only a randomly selected subset of the applicants would actually receive the assistance. The potential downside of this approach is limited external validity in the sense that the estimated treatment effect will apply only to the self-selected group of applicants – unless, of course, that is the impact of interest to the evaluators and policy-makers (McKenzie, 2010).

Alternatively, the treatment itself could be defined in a way that is de facto compulsory so the question of selective take-up does not arise. Thus, a randomly selected set of firms would receive some form of “encouragement,” for example through phone calls or visits aimed at providing detailed information on the application process, raising the probability that those firms apply to the program.⁴⁵ The unbiased effect of this random encouragement - the “intention-to-treat” effect - would be estimated by comparing take-up by firms that received encouragement relative to take-up by firms that did not. To obtain the effect of the program on ultimate outcomes (e.g., export performance) one would instrument for treatment using the randomly provided encouragement. A limitation of this approach is that in the presence of “essential” heterogeneity”, out of the set of firms receiving the encouragement, those that take up the program are likely to have higher unobserved expected benefits from the program than those that

⁴⁵ Duflo, Kremer, and Robinson (2006) offer an example of an encouragement design. They tested whether seeing a neighbor use fertilizers would encourage other farmers to do the same. For each using farmer, they invited randomly chosen neighbors to attend a demonstration of fertilizer use. Although other farmers were also welcome to attend, the attendance rate was much higher in the sub-sample of invited ones, which was randomized. To our knowledge, no trade intervention has been evaluated with encouragement design.

do not. Hence encouragement design could be associated with biased treatment effects, which are potentially over-estimated relative to the average effects for the sample that would have taken up treatment in the absence of encouragement (Barrett and Carter, 2010; McKenzie, 2010).

7. Conclusion

In spite of the challenges, rising demands for results and accountability from donors and clients alike require that aid-for-trade evaluation strategies need more ambition and rigor. Implementing agencies should no longer be content with traditional methods based on output monitoring and before-after comparisons. Output monitoring is largely introspective, relying on measures defined by the task managers and therefore liable to biases, while before-after comparisons are vulnerable to confounding influences.

The basic problem faced in the evaluation of a policy, program or project impact is *attribution*. Are the observed changes in the performance of treated entities really attributable to the intervention under consideration, or do they reflect a fortuitous combination of effects? Impact-evaluation (IE) methods — developed outside of the social sciences but widely adopted in the evaluation of poverty, health and education programs — provide a generally accepted answer to the problem of attribution.

Trade interventions have so far escaped the rising tide of evaluation methods. But there is no justification for this trade exceptionalism as IE techniques are many and sufficiently flexible for use even in the case of interventions that are not targeted at a defined group of treated individuals.

As the authors have experienced in their campaign for greater recourse to IE techniques in trade, the key barriers to progress are not conceptual. Rather, they concern incentive issues, as IEs are costly, burdensome, lengthy, and not necessarily aligned with project managers' incentives. For example, World Bank projects to assist private sector firms in Africa last on average five years, which would imply that, if their IE involved an RCT, many years would need to elapse for the projects to show results (McKenzie, 2011a). These many years would go well beyond a project manager's horizon.

In principle, researchers need not wait until completion of the project to evaluate its effects; rather, results one or two years after the project could be assessed and be used to guide the implementation of the project in the subsequent years. While early feedback from an IE is useful, is useful it should however be treated with caution. First, it may simply be premature in that the effects of the program may not be adequately manifested. Second, from a methodological perspective, fine-tuning a program at an intermediate stage could jeopardize the possibility of evaluating its effects credibly.

The weakness of current evaluation practice can be illustrated no better than by this critical assessment, found in the Implementation Completion Report of a recent World Bank project in the area of export promotion:

Although the design of the M&E system was appropriate, both Bank and Government project teams had difficulty measuring the achievements of the project using the broad indicators cited in the PAD.⁴⁶ [B]y current standards, they were insufficient and incomplete.

[...] M&E, particularly important as a learning objective, was weak. It was slow to start and did not deliver. The M&E staff [...] lacked the capacity and experience to carry out the monitoring activities, and the Unit was unable to carry out baseline and impact surveys of randomly selected farmers in both project and non-project areas, i.e., survey to gauge key interest groups' response to the outputs generated by the pilot activities. The M&E Unit's ability to collaborate with other implementing agencies to collect information and data was also ineffective. Implementing partners did not regard the M&E exercise as a learning process but instead, conducted their promotion activities without consulting or collaborating with the M&E unit.

As the reviewers noted, the learning function of evaluation tends to be overshadowed by the “monitoring” function for implementing agencies.

In order to overcome these hurdles, several avenues must be considered. First, the burden imposed on project managers should be relieved by making impact evaluation a separate exercise carried out by specialists, albeit in collaboration with project managers. Project managers should be involved at the right time — that is, during project design and from then on, as much as possible, left in peace. The World Bank has moved in this direction through the creation of the DIME unit, which provides expertise and help with IE financing.

At the same time, governments in the countries receiving trade assistance must buy into the process. This means sharing knowledge and building capacities for a proper interpretation of IE results and, over the long run, for governments to build their own IE capabilities as part of public-services delivery improvements.

Also, every effort should be made to reduce the cost of IEs. For small-scale activities, the cost of an IE can be as great as that of the activity itself. This is excessive. Local resources — in particular universities and graduate students — should be involved, producing a double benefit: costs are reduced and local capacities are strengthened.

Finally, the exploitation of IE results should prioritize learning over monitoring. That is, donors and implementing agencies should tread cautiously in using IE results to frame incentive systems. Care is needed in the interpretation of IE results because premature conclusions could

⁴⁶ M&E stands for Monitoring and Evaluation while PAD stands for Project Appraisal Document.

easily provoke a backlash and because a considerable accumulation of evidence is needed to yield truly valuable new knowledge.

References

Abbring, J. and J. Heckman (2007). “Econometric Evaluation of Social Programs Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation,” in Heckman, J. and E. Leamer (eds.) *Handbook of Econometrics*, vol. 6B, pp. 5146-5303.

Ashenfelter, O., Harmon, C., and H. Oosterbeek (1999). “A Review of Estimates of the Schooling/Earnings Relationship,” *Labour Economics* 6, 453-470.

Atkin, D. and A. Khandelwal (2011). “The Use of Experimental Designs in the Evaluation of Trade-Facilitation Programs,” in Cadot, O., Fernandes, A., Gourdon, J., and A. Mattoo (eds.) *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, pp. 107-122. The World Bank and CEPR.

Balat, J., Brambilla, I., and G. Porto (2009). “Realizing the Gains from Trade: Export Crops, Marketing Costs, and Poverty,” *Journal of International Economics* 78, 21-31.

- Banerjee, A., S. Jacob, M. Kremer, J. Lanjouw, and P. Lanjouw (2005). “Moving to Universal Education: Costs and Trade-offs,” MIT mimeo.
- Banerjee, A., Amsden, A., Bates, R., and J. Bhagwati, and N. Stern (2007). *Making Aid Work*. MIT Press.
- Banerjee, A. and E. Duflo (2008). “The Experimental Approach to Development Economics,” NBER Working Paper 14467.
- Barrett, C. and M. Carter (2010). “The power and Pitfalls of Experiments in Development Economics: Some Non-Random Reflections,” *Applied Economic Perspectives and Policy* 32, 515-548.
- Blundell, R. and M. Costa Dias (2009). “Alternative Approaches to Evaluation in Empirical Microeconomics,” *Journal of Human Resources* 44, 565-640.
- Brenton, P. and E. von Uexkuhl (2009). “Product-Specific Technical Assistance for Exports—Has it Been Effective?,” *Journal of International Trade and Economic Development* 18, 235-254.
- Bruhn, M. (2011). “License to Sell: The Effect of Business Registration Reform on Entrepreneurial Activity in Mexico,” *Review of Economics and Statistics* 93, 382-386.
- Cali, M. and D. te Velde (2011). “Does Aid for Trade Really Improve Trade Performance?,” *World Development* 39, 725-740.
- Caliendo, M. and S. Kopeinig (2005). “Some Practical Guidance for the Implementation of Propensity Score Matching,” IZA Discussion Paper 1588.
- Campbell, D. (1969). “Reforms as Experiments,” *American Psychologist* 24, 407-429.
- Cantens, T., Raballand, G., Bilangna, S., and M. Djeuwo (2011). “Reforming Customs by Measuring Performance: a Cameroon Case Study,” in Cadot, O., Fernandes, A., Gourdon, J. and A. Mattoo (eds.) *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, pp. 183-206. The World Bank and CEPR.
- Cook, T., Shadish, W., and V. Wong (2006). “Within Study Comparisons of Experiments and Non-Experiments: Can they help decide on Evaluation Policy?,” Northwestern University mimeo.
- Datt, M. and D. Yang (2011). “Half-Baked Interventions: Staggered Pre-shipment Inspections in the Philippines and Colombia,” in Cadot, O., Fernandes, A., Gourdon, J. and A. Mattoo (eds.) *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, pp. 163-182. The World Bank and CEPR.
- Djankov, S., Freund, C. and C. Pham (2010). “Trading on Time,” *Review of Economics and Statistics* 92, 166-173.

- Djebbari, H. and J. Smith (2008). "Heterogeneous Program Impacts of PROGRESA," *Journal of Econometrics* 145, 64-80.
- Duflo, E., Kremer, M., and J. Robinson (2006). "Understanding Technology Adoption: Fertilizer in Western Kenya, Preliminary Results from Field Experiments," Mimeo, MIT.
- Duflo, E., Glennerster, R., and M. Kremer (2008). "Using Randomization in Development Economics Research: A Toolkit," in Schultz, T.P. and J. Strauss (Eds.) *Handbook of Development Economics*, vol. 4, pp. 3895-3962.
- Fan, Y. and S. Park (2010) "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference," *Econometric Theory* 26, 931-951.
- Ferro, E., Portugal-Perez, A., and J. Wilson (2011). "Aid-for-Trade and Export Performance: The Case of Aid in Services," in Cadot, O., Fernandes, A., Gourdon, J. and A. Mattoo (eds.) *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, pp. 207-219. The World Bank and CEPR.
- Francois, J. and M. Manchin (2007). "Institutions, Infrastructure, and Trade," Policy Research Working Paper Series 4152.
- Freund, C. and M. Pierola (2010). "Export Entrepreneurs: Evidence from Peru," World Bank Policy Research Working Paper 5407.
- Freund, C. and M. Pierola (2011). "Export Superstars," World Bank mimeo.
- Freund, C. and N. Rocha (2011). "What Constrains Africa's Exports," *World Bank Economic Review* 25, 361-386.
- Gamberoni, E. and R. Newfarmer (2009). "Aid for Trade: Matching Potential Demand and Supply," World Bank Policy Research Working Paper 4991.
- Gine, X. and I. Love (2011). "Do Reorganization Costs Matter for Efficiency? Evidence from a Bankruptcy Reform in Colombia," *Journal of Law and Economics*, forthcoming.
- Glazerman, S., Levy, D., and D. Myers (2003). *Nonexperimental Replications of Social Experiments: A Systematic Review*. Princeton, NJ: Mathematica Policy Research, Inc.
- Gourdon, J., Marchat, J., Sharma, S. and T. Vishwanath (2011). "Can Matching Grants Promote Exports? Evidence from Tunisia's FAMEX Program," in Cadot, O., Fernandes, A., Gourdon, J. and A. Mattoo (eds.) *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, pp. 81-106. The World Bank and CEPR.
- Harrison, A. and A. Rodríguez-Clare (2010). "Trade, Foreign Investment, and Industrial Policy," in Rodrik, D. and M. Rosenzweig (eds.) *Handbook of Development Economics* vol. 5 pp. 4039-4214.

Heckman, J., Urzua, S., and E. Vytlacil (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics* 88, 389-432.

Helble, M., Mann, C. and J. Wilson (2009). "Aid for Trade Facilitation," World Bank Policy Research Working Paper 5064.

Hoekman, B. and A. Nicita (2008). "Trade Policy, Trade Costs, and Developing Country Trade," World Bank Policy Research Working Paper 4797.

Hausmann, R., Hwang, J., and D. Rodrik (2007). "What You Export Matters," *Journal of Economic Growth* 12, 1-25.

IEG (2006). *Assessing World Bank Support for Trade, 1987-2004: an IEG Evaluation*. The World Bank.

Imbens, G. and J. Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature* 47, 5-86.

Jaud, M. and O. Cadot (2011). "A Second Look at the Pesticides Initiative Program: Evidence from Senegal," World Bank Policy Research Working Paper 5635.

Khandker, S., Koolwal, G., and H. Samad (2010). *Handbook on Impact Evaluation*. Washington DC: The World Bank.

Klapper, L. and I. Love (2010). "The Impact of Business Environment Reforms on New Firm Registration," World Bank Policy Research Working Paper 5493.

Lalonde, R. (1986). "Evaluating the Econometric Evaluations of Training Programs Using Experimental Data," *American Economic Review* 76, 602-620.

Lederman, D., M. Olarreaga, and L. Payton (2010). "Export Promotion Agencies Revisited," *Journal of Development Economics* 91, 257-265.

Lederman, D., Rodríguez-Clare, A., and D. Xu (2011). "Entrepreneurship and the extensive margin in export growth: a microeconomic accounting of Costa Rica's export growth during 1997-2007," *World Bank Economic Review* 25, 543-561.

Lopez-Acevedo, G. and M. Tinajero (2010). "Mexico: Impact Evaluation of SME Programs using Panel Firm Data," World Bank Policy Research Working Paper 5186.

Manski, C. (1997). "The Mixing Problem in Programme Evaluation," *Review of Economic Studies* 64, 537-553.

McKenzie, D. (2010). "Impact Assessments in Finance and Private-Sector Development: What Have We Learned and What Should We Learn?" *World Bank Research Observer* 25, 209-233.

- McKenzie, D. (2011a). “How Can We Learn Whether Firm Policies Are Working in Africa? Challenges (and Solutions?) for Experiments and Structural Models,” World Bank Policy Research Working Paper 5632.
- McKenzie, D. (2011b) “Beyond Baseline and Follow-up: The Case for More T in Experiments,” World Bank Policy Research Working Paper 5639.
- Miguel, E. and M. Kremer (2004). “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities,” *Econometrica* 72, 159-217.
- Morduch, J. (1998). “Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh,” Princeton University, Woodrow Wilson School of Public and International Affairs, Research Program in Development Studies Working Paper 198.
- Nelson, D. and S. Silva (2008). “Does Aid Cause Trade? Evidence from an Asymmetric Gravity Model,” University of Nottingham research paper 2008/21.
- Osei, R., O. Morrissey, and T. Lloyd (2004). “The Nature of Aid and Trade Relationships,” *European Journal of Development Research* 16, 354-374.
- Portugal-Perez, A. and J. Wilson (2010). “Export Performance and Trade Facilitation Reform: Hard and Soft Infrastructure,” World Bank Policy Research Working Paper 5261.
- Rajan, R. and A. Subramanian (2008). “Aid and Growth: What Does the Cross-Country Evidence Really Show?,” *Review of Economics and Statistics* 90, 643-665.
- Ravallion, M. (2008). “Evaluating Anti-Poverty Programs,” in Schultz, T.P. and J. Strauss (eds.) *Handbook of Development Economics*, vol. 4 pp. 3787-3846.
- Ravallion, M. (2009). “Evaluation in the Practice of Development” *World Bank Economic Observer* 24, 29-53.
- Ravallion, M. (2011). “On the Implications of Essential Heterogeneity for Estimating Causal Impacts using Social Experiments,” World Bank Policy Research Working Paper [5804](#).
- Rodrik, D. (2006). “What's So Special about China's Exports?,” *China and World Economy* 14, 1-19.
- Rodrik, D. (2008). “The New Development Economics: We Shall Experiment, but Shall We Learn?,” Mimeo, John F. Kennedy School of Government, Harvard University.
- Rosenbaum, P. and D. Rubin (1983). “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika* 70, 41-55.
- Savedoff, W. (2006). *The Evaluation Gap: An International Initiative to Build Knowledge*. Center for Global Development, Washington, D.C..

Sequeira, S. (2011). "Transport Costs and Firm Behavior," in Cadot, O., Fernandes, A., Gourdon, J., and A. Mattoo (eds.) *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, pp. 123-162. The World Bank and CEPR.

Tan, H. (2009). "Evaluating SME Support Programs in Chile using Panel Firm Data," World Bank Policy Research Working Paper 5082.

Todd, P. (2008). "Evaluating Social Programs with Endogenous Program Placement and Self Selection of the Treated," in Schultz, T.P. and J. Strauss (eds.) *Handbook of Development Economics*, vol. 4, pp. 3848-3894.

Volpe, C. and J. Carballo (2010). "Beyond the Average Effects: The Distributional Impacts of Export Promotion Programs in Developing Countries," *Journal of Development Economics* 92, 201-214.

Volpe, C. (2011). "Assessing the Impacts of Trade Promotion Interventions: Where Do We Stand?," in Cadot, O., Fernandes, A., Gourdon, J., and A. Mattoo (eds.) *Where to Spend the Next Million? Applying Impact Evaluation to Trade Assistance*, pp. 39-80. The World Bank and CEPR.

Wagner, D. (2003). "Aid and Trade - an Empirical Study," *Journal of the Japanese and International Economies* 17, 153-173.

WTO (2009). *World Trade Report 2009: Trade Policy Commitments and Contingency Measures*. Geneva: World Trade Organization.

World Bank (2009). *Unlocking Global Opportunities: The Aid For Trade Program of the World Bank Group*. Washington, DC: The World Bank.

World Bank (2011). "Leveraging Trade for Development and Growth: The World Bank Group Trade Strategy, 2011-2021," Washington, DC: The World Bank.