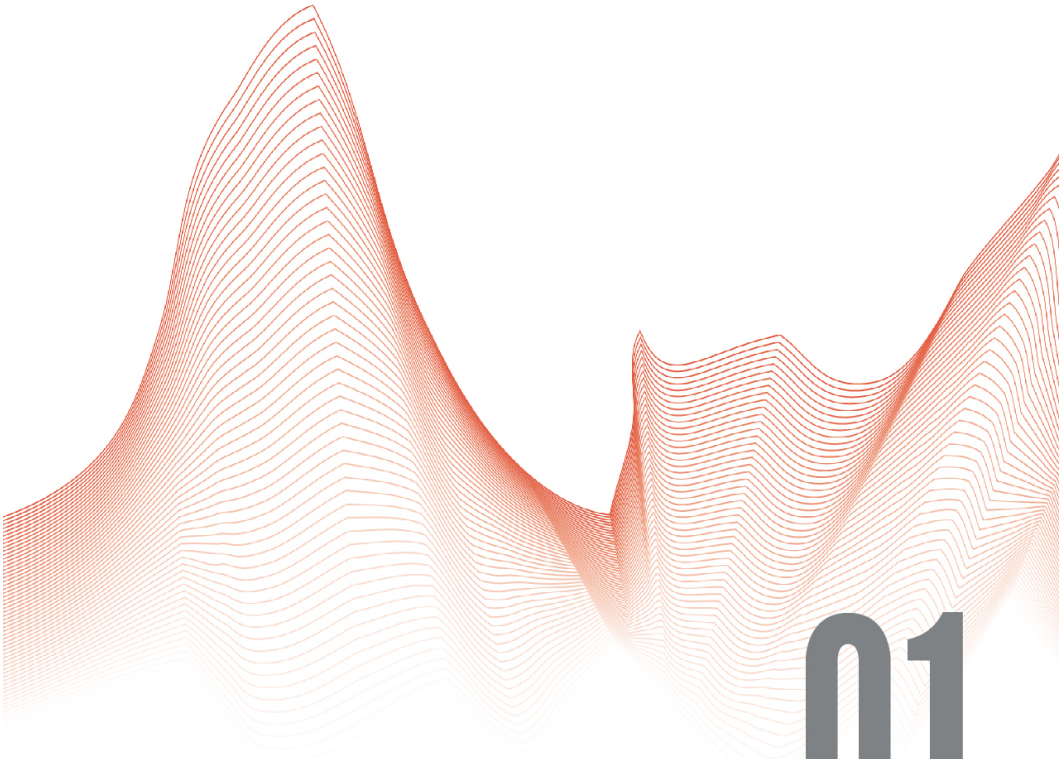
EBA



01
2014

RANDOMIZED CONTROLLED TRIALS: STRENGTHS, WEAKNESSES AND POLICY RELEVANCE

Anders Olofsgård

Randomized Controlled Trials:

Strengths, Weaknesses and Policy Relevance

Anders Olofsgård

Rapport 2014:1
till
Expertgruppen för biståndsanalys (EBA)

This report can be downloaded free of charge at www.eba.se.
Hard copies are on sale at Fritzes Customer Service.

Address:
Fritzes, Customer Service,
SE-106 47 Stockholm
Sweden
Fax: 08 598 191 91 (national)
+46 8 598 191 91 (international)
Tel: 08 598 191 90 (national)
+46 8 598 191 90 (international)
E-mail: order.fritzes@nj.se
Internet: www.fritzes.se

Printed by Elanders Sverige AB
Stockholm 2014

Cover design by Julia Demchenko
ISBN 978-91-38-24114-1

Förord

Regeringens huvudsakliga resultatstyrning av utvecklingsbiståndet sker genom så kallade resultatstrategier. Av regeringens riktlinjer för dessa strategier framgår att ”beslut om biståndets fortsatta utformning i hög grad ska fattas utifrån en analys av de resultat som har uppnåtts.”[1] Resultatredovisningen blir därmed styrande för biståndets framtida utformning. Det ställer i sin tur höga krav på hur förväntade resultat uttrycks och på mätning och rapportering av resultat.

Resultat kan uttryckas som prestationer eller effekter. Prestationer (”output” på engelska) är varor eller tjänster som följer som ett direkt resultat av genomförda utvecklingsinsatser. Sådana resultat är i allmänhet lätta att mäta och rapportera (till exempel ett anordnat hållbarhetsseminarium, en anlagd väg eller antal vaccinationer). Effekter (”outcome” eller ”impact” på engelska) är resultat som orsakas av prestationerna, det vill säga den del i ett förändrat tillstånd (”hållbarhet”, bättre fungerande marknader, eller en utrotad sjukdom) som beror på utvecklingsinsatsen. Effekter är i allmänhet svårare att mäta bland annat på grund av att den observerade tillstånds-förändringen nästan aldrig beror enbart på prestationerna. Av regeringens riktlinjer för resultatstrategier framgår att resultat uttryckta i form av effekter eftersträvas men att resultat i form av prestationer kan formuleras om förväntade effekter inte är möjliga att uttrycka.

[1] Riktlinjer för resultatstrategier inom Sveriges internationella bistånd, Promemoria 2013-07-11, UF2013/41712/UD/USTYR, sid. 2.

Regeringens resultatstrategier för biståndsverksamhet i länder är korta dokument inriktade på att beskriva vilka resultat den svenska biståndsverksamheten förväntas leda till under strategiperioden. Till exempel innehåller resultatstrategin för Tanzania 2013-2019 bland annat följande förväntade resultat: utvecklade marknader i jordbruksproduktionen; ökad rättssäkerhet kring landrättigheter; ett ökat antal elever som tillgodogör sig grundläggande kunskaper och färdigheter i skolan; ett ökat antal unga som genomgår yrkesutbildning; ökade möjligheter för kvinnor och unga att starta och driva produktiva företag; minskad korruption inom offentlig förvaltning och stärkt kapacitet inom det civila samhället att utkräva ansvar. För att tydligt kunna påvisa att resultat av ovanstående karaktär kan tillskrivas insatser finansierade med svenska biståndsmedel krävs i regel så kallade effektutvärderingar.

Mot denna bakgrund gav EBA docent Anders Olofsgård, verksam vid SITE, Handelshögskolan i Stockholm, i uppdrag att göra en litteraturstudie av en ledande metod för effektutvärdering, randomiserade kontrollstudier. Studien avser beskriva metoden samt belysa när det är önskvärt och under vilka förutsättningar det är möjligt att använda denna metod för utvärdering inom internationellt utvecklingssamarbete.

Författaren identifierar två olika användningsområden för randomiserade kontrollstudier inom biståndet. En forskningsnära användning är att studera beteendeeffekter av generella typer av insatser i syfte att styra den övergripande planeringen av framtida biståndsverksamhet. En mer direkt användning är att utvärdera enskilda projekt, för att besluta om liknande projekt eller om fortsatt

finansiering eller avslut. Metoden har styrkor och svagheter, och den relativa betydelsen av dessa beror bland annat på vilket användningsområde som avses. Ytterligare insikter kan nås genom att kombinera metoden med andra, till exempel kvalitativa, metoder (så kallade ”mixed methods”) samtidigt som randomiserade kontrollstudier under rätt omständigheter är den bästa metod som finns praktiskt tillgänglig för effektutvärdering.

Författaren hävdar också att en mer systematisk användning av randomiserade kontrollstudier har fördelar bortom en förståelse av biståndsinsatsers effekter. Eftersom metoden inkluderar gedigna förstudier och kräver tydliga målformuleringar kan den bidra till att stärka biståndsverksamheten i stort kvalitetsmässigt. Sveriges trovärdighet som biståndsaktör kan också påverkas positivt genom att metoder används, som anses vara av högsta kvalitet och som i ökande utsträckning tillämpas av andra ledande givare.

Författaren menar vidare att befintlig evidens på området (tidigare effektutvärderingar utförda av andra) systematiskt bör tillgängliggöras för handläggande personal och att dessa ges kompetens att bedöma tillförlitlighet i utvärderingar och att värdera styrkan av presenterade resultat. Randomiserade kontrollstudier i egen (Sidas) regi bör också övervägas.

Resultatstrategierna beslutades i och med budgetpropositionen för år 2013 och ersätter successivt tidigare strategier för biståndsverksamhet i länder och regioner, tematiskt inriktat bistånd samt bistånd till multilaterala organisationer. Styrformen är således ny och huruvida resultatrapporteringskravet uppfylls går således inte att bedöma ännu.

Det är min förhoppning att denna rapport ska utgöra ett bra underlag för den fortsatta diskussionen om effektutvärdering av svenskfinansierad biståndsverksamhet. Arbetet med rapporten har följts av en referensgrupp under ledning av Professor Jakob Svensson, ledamot i EBA:s styrelse. Författaren ansvarar själv för innehållet i rapporten.

Stockholm i maj 2014

Lars Heikensten

Ordförande i EBA

Table of contents

Sammanfattning

Det övergripande syftet med denna rapport är att bidra till den pågående debatten om utvärdering och effektivitet i svenskt bistånd. Mer specifikt ligger fokus på betydelsen av så kallade randomiserade kontrollerade studier (RKS) i utvärderingen av biståndsfinansierade projekt och program. Den första delen av rapporten behandlar hur RKS har använts i den internationella akademiska världen samt för att främja evidensbaserad utvecklingspolitik. Den andra delen diskuterar i vilken utsträckning denna metod använts inom svensk biståndsutövning och om det finns skäl att utvidga användningen. Jag föreslår också slutligen vad som kan göras mer konkret i Sverige för att bättre ta tillvara de möjligheter till kunskap som metoden ger.

Biståndsutvärdering kan delas upp i processutvärdering, med fokus på design och implementering samt de omedelbara resultaten av ett projekt eller program, och effektutvärdering, med fokus på effekterna av projektet eller programmet på den definierade målgruppen. Att mäta effekt har generellt setts som svårare än att mäta resultat. Jämför utmaningen i att mäta utfallet av ett skolbyggnadsprojekt i termer av ökad kunskap bland barnen i det aktuella området (effekt) med att mäta antalet skolor byggda (resultat, eller prestation). Framförallt har det varit svårt att säkerställa att observerade utfall verkligen är kausala effekter (orsakade av projektet eller programmet) och inte bara korrelationer (observerade samtidiga förändringar i utfallsmått med projektet eller programmet, eventuellt helt utan orsakssamband). RKS erbjuder en lösning på detta problem, och har därför haft ett stort inflytande på utvecklingsekonomisk forskning, men metoden blir också alltmer betydelsefull inom den utvecklingspolitiska debatt som

förs mellan regeringar, biståndsorgan och internationella finansiella institutioner.

RKS är under de rätta omständigheterna den bästa metod som finns praktiskt tillgänglig för effektutvärdering och kan hjälpa beslutsfattare att bättre allokera resurser så att biståndet får en reell effekt. Metoden bör därför vara en del av ledande biståndsorganisationers utvärderingsportfölj. Metoden har också betydande begränsningar, den kan inte besvara alla relevanta frågor och kan inte tillämpas på alla typer av projekt och program. Den kan alltså aldrig vara mer än ett av många verktyg. Det idealiska tillvägagångsättet är ofta att använda flera kompletterande utvärderingsmetoder (”mixed methods”), dvs. både RKS och andra kvantitativa och kvalitativa metoder. Vad gäller svensk praktik är det dock så att RKS har använts så sparsamt att det finns ett stort utrymme för att öka kvaliteten på effektutvärderingar inom biståndshanteringen.

Ett mer systematiskt användande av RKS inom den svenska biståndsverksamheten har flera fördelar utöver en bättre förståelse av biståndets effekter. För det första skulle det stärka Sveriges trovärdighet som en seriös och ledande partner inom den internationella biståndspolitiken i och med att denna metod används alltmer av andra ledande bilaterala och multilaterala givare. För det andra kräver metoden en tydlig definition av vad målen med en insats är och hur måluppfyllelse förväntas mätas och bedömas redan vid initieringsfasen. Detta kan vara en stor hjälp för att undvika vanliga fallgropar inom biståndsverksamheten; oklara mål, förväntade effekter

som inte kan mätas eller kvantifieras, samt bristande förstudier som gör det omöjligt att bedöma ifall förbättringar skett eller inte.

Det finns flera insatser som kan göras för att främja utvecklingen mot en mer evidensbaserad planering av biståndsfinansierade projekt och program. För det första kan de befintliga resurser som finns i termer av effektutvärderingar utförda av andra användas mer systematiskt. Detta kräver att personal som beslutar om finansiering av projekt och program har tillgång till befintligt material, vet hur man läser och bedömer olika typer av utvärderingar, samt kan värdera tillförlitligheten och begränsningarna (såsom extern validitet) i olika utvärderingsmetoder. Detta kan kräva fortbildning och tillgång till rådgivning och hjälp från personal specialiserade på just utvärderingsfrågor. Ett lovvärt initiativ för att ta till sig mer av den befintliga kunskapen är forskningsprogrammet ReCom, samfinansierat av Sida och Danida.

Ett andra steg som bör tas är att stärka kompetensen att upphandla rigorösa effektutvärderingar inom Sida. Slutligen, för att utföra RKS så krävs ett samarbete mellan finansiären (vanligen Sida) och representanter för partnerländerna och de organisationer eller företag som genomför själva interventionen som ska utvärderas. Detta kan i sig kräva gemensamma utbildningsinsatser om varför metoden är önskvärd och vilka krav den ställer på de olika aktörernas agerande i olika faser av genomförandet. Det finns också en möjlighet att kombinera direkt biståndsverksamhet med utvecklandet av en resursbas för effektutvärdering genom att stödja kompetensuppbyggnad kring hur man genomför effektutvärderingar på universitet och forskningscentrum inom mottagarländerna.

Executive summary

The general purpose of this report is to offer an input into the ongoing debate on evaluation and effectiveness of Swedish foreign aid. More specifically, the focus lies on the role of so called Randomized Controlled Trials (RCTs) in evaluations of impact of aid financed activities. The first part brings up how RCTs have been used in research and practice to promote evidence based development policy. The second part discusses to what extent, up until now, this tool has been used by Swedish aid practitioners, if there are reasons to expand that use, and finally offers some tentative suggestion for how that can be achieved.

Aid evaluation entails both process evaluation, focusing on design and implementation and the immediate output of the project/program, and impact evaluation, focusing on the impact of the project/program on the defined beneficiaries. Measuring impact has generally been seen as more difficult than measuring output, in particular with regards to establishing a causal impact rather than mere correlation. The RCT methodology offers a solution to this problem, and has had a big influence on the academic field of development economics, but is also gaining influence in the debate on development policy within governments, aid agencies and international financial institutions.

In this report I argue that RCTs is a powerful and high quality method to evaluate impact under the correct circumstances, and can help decision makers better allocate resources towards interventions that make a real difference in the life of aid recipients. It should

therefore be part of the toolbox of aid agencies that have an obligation to make sure that development finance is allocated towards interventions that work. RCTs also have significant limitations, though, and can neither generate answers to all policy relevant questions, nor be applied to all types of projects and programs. It can thus be no more than one of many tools for monitoring and evaluation used by aid agencies. Using mixed methods, a combination of RCTs and other quantitative and qualitative methods, will often be the ideal approach to learn more broadly about the effectiveness of different dimensions of aid financed interventions. Nevertheless, the fact that RCTs have been used very sparsely, if at all, by the Swedish aid community suggests that there is a lot of potential to increase the quality of impact evaluation of projects and programs financed by Swedish aid.

I also argue that the advantages of including RCTs into Swedish aid practices go beyond just getting a better understanding of impact. First, it would lend more credibility to the ambition to be a serious and leading partner in the international aid community as this is a tool increasingly used by other bilateral and multilateral donors. Second, RCTs require a firm understanding of what exactly the objectives of the intervention are, and how their fulfilment can be measured, already from the start. This helps avoid common pitfalls with aid financed interventions; unclear objectives, unobservable or unmeasurable intended outcomes, and the inability to even quantify changes in outcomes in the targeted group due to missing baselines.

To move forward towards a more evidence based planning of projects and programs, several steps can be made concurrently. First,

use more systematically the existing bank of knowledge in the form of already done rigorous impact evaluations. This requires that staff have access to existing material, know how to read and evaluate evidence derived using different methods, and know how to address limitations with regards to for instance external validity. This may require training and access to advice and help from staff specialized in monitoring and evaluation. The ReCom initiative is a good first step in that direction. Second, acquire the in-house competence to commission RCTs of projects and programs financed by Sida or collaborators. Finally, conducting RCTs require the collaboration of partner countries and implementing units (NGOs or consultancy firms). This may require collaborative training efforts, and at times some convincing. A possibly fruitful approach to combine partner country human capital development with the creation of a resource for conducting RCTs is to support development of rigorous impact evaluation skills at partner country universities and research centres.

Introduction

This report is written in a context where the organizational task of aid evaluation in Sweden is in flux. SADEV has been terminated after a critical report from Statskontoret (2012) and the evaluation unit at Sida has been reorganized within the organization. The upside of this is that it has created a political momentum to think through how this important task can be strengthened going forward. The purpose of this report is to contribute to that mission by discussing the role of so called Randomized Controlled Trials (RCT) in evaluating aid financed projects and programs.

As within other public policy areas, evaluation of programs and projects are essential for learning what works and how to design and operate interventions. There are some aspects that make this particularly challenging in the context of foreign aid, such as the foreign environment, the multiplicity of stakeholders, and the very broad range of types of projects and programs.[1] Nevertheless, or maybe just because of that fact, there is typically no shortage in terms of the quantity of aid evaluations, even leading some observers to complain about an "obsessive measurement disorder" (Natsios, 2010). Equally important, though, is the question of the quality of the evaluations, and the balance between the objectives of the evaluations.

[1] Foreign aid interventions span all areas of public policy; health and education, infrastructure, financial policy, legal reforms, etc. This means that decision makers need to have a very broad capacity to commission, read and judge evaluations across quite different fields, potentially using different methodologies suited for the specific question. A general challenge is also to attribute the effect of Swedish aid generally in a context where many other donors are involved and resources are fungible.

Evaluations within foreign aid are supposed to deliver a systematic and objective assessment of the design, implementation and results of projects and programs (OECD, 2002). An often made distinction is that between process evaluation and impact evaluation. To make the distinction concrete, think of a school construction program aiming at better access to schools and more and better educated students. Process evaluation is typically focusing on design and implementation, and is primarily concerned with the output of the project/program. In our example, process evaluation would focus on how many schools were built (in numbers and in relation to the objective) and how success/failure in achieving the building targets depended on details of how the project/program was planned, operated and supervised. Impact evaluation instead focuses on the impact of the project/program on the defined beneficiaries. In our example; did the new schools also contribute to increased school enrolment, better test scores, less teacher absenteeism, and so on? Methodologically, though, counting the number of schools is substantially easier than to establish a causal effect on test scores. Furthermore, incentives within aid agencies are typically more directed towards making sure that the allocated budget is spent smoothly and without financial leakage than to guarantee impact (e.g. Martens 2002). Actual aid evaluation is therefore predominantly of the process evaluation type, and rigorous methods to get at impact are rarely used. Not denying the importance of process evaluation, the difference between output and impact is important, and it is important to have balance between the two types of evaluation.

RCTs are explicitly designed to evaluate causal impact. The application of the methodology has become a major field among development economists in academia. The general method and terminology are not new but borrowed from the medical sciences and drug studies in particular. What is relatively new is the use of this in economics, and there are probably few areas in which it has been so enthusiastically applied as in development economics. In short (discussed more in detail in the next section), the methodology requires a sample of the unit under study, for instance individuals, households, schools, or health clinics, a well-defined intervention with observable and measurable outcomes, and a random selection of the units into treated (those getting the intervention) and control (those not getting the intervention). The process typically starts with a baseline survey to establish that the groups of treated and control share similar characteristics. Then the actual intervention is implemented. Finally, after a sufficient amount of time to let potential effects develop, an end-line survey is taken that (typically) together with the baseline is used to derive estimates of the intervention's actual impact.

The popularity of the approach largely stems from that it purportedly offers a solution to a common problem in evaluation; how to distinguish causal impact from mere correlation. For example, a project that offers all teachers in a school district flip-charts to use in their teaching could possibly be evaluated based on improvements in test scores in classes where teachers picked up on the new pedagogical tool relative to the improvement in classes where teachers did not. However, a reasonable suspicion could be that teachers more

determined to improve test scores in their classes are more likely to test the new technology. The impact of the flip-charts are then confounded by the difference in ambition, and all other behavioural changes that may come with that, making it hard to separate the causal effect on test scores of the flip-charts from that of simply having more ambitious teachers. Randomly assigning flip-charts across schools can eliminate, or at least alleviate, that problem. However, the approach also has important limitations in different ways, not the least for guiding policy. There is a valid concern that the portfolio of interventions becomes biased towards what can be randomized rather than driven by needs and ex ante estimated relevance, and that aimed for impact becomes too short run at the cost of long run institutional development harder to measure and quantify. These, and other limitations, will be discussed more in detail in Section 3.

Given the centrality of this approach in development research and its very direct and concrete link to evaluation of often at least partially aid financed interventions, a good understanding of its strengths and weaknesses and how and when to use it in a policy context seems central. The purpose of this report is thus to offer an overview of the debate around the use of the methodology and in particular the trade-offs involved when thinking about it as a tool to guide policy. Furthermore, we will also discuss what would be required within the Swedish aid context to use the approach more systematically, and why this would be beneficial. As will be discussed, benefits may go beyond just getting a better understanding of impact, but it may also require some re-orientation of how aid is currently organized.

The RCT method and development applications

The modern randomized control trial is typically associated with the medical sciences (though examples of randomized experiments date back to the 19th century, for instance in the areas of psychology and education). The applications to international development have largely borrowed that methodology and terminology with a few exceptions. For instance, in clinical studies, the group that doesn't get the substance, referred to as the control group (the existence of which is a distinct feature of the RCT), typically gets a placebo or an alternative pre-existing drug, and subjects do not know if they belong to the treatment or control group. This "blind" treatment is typically not possible with the type of interventions done outside of clinical trials. Subjects know if they are part of the treatment group or not. It is therefore generally a bit more difficult to attribute the effects to a very specific mechanism, and it is important to understand that the control group may react negatively to not being selected to receive the treatment.[2] To reduce the negative reaction among the control group, the intervention is often rolled out in sequence, so today's control group will be tomorrows treatment group (once the evaluation of the initial intervention is done). On the other hand, in many development interventions there is also a value to understanding the behavioural response to being selected as beneficiaries, as this is a likely feature of

[2] To better identify the mechanism, it has become popular to use so called mixed methods, combining a randomized impact evaluation with for instance field experiments playing simple games to elicit the impact of the intervention on trust and cooperation (e.g. Desai, Joshi and Olofsgard, 2014).

any program, so not having double blind interventions is not necessarily a disadvantage.

Why is the selection into treatment and control random? The purpose of impact evaluation is to establish the causal effect of a program at hand. Strictly speaking this requires an answer to a counterfactual question; what difference does it make for the average individual if he is part of the program or not. Since an individual cannot be both part of and not part of the program at the same time, an exact answer to that question cannot be reached. Instead evaluators must rely on a comparison between individuals participating in the program and those that do not, or a before and after comparison of program participants. The challenge when doing this is to avoid getting the comparison contaminated by unobservable confounding factors and selection bias. For instance, maybe only the already most motivated households are willing to sign up for a conditional cash transfer program offering cash in exchange for school attendance. In this case an observational study finding a positive correlation between program involvement and school participation may all be due to a selection bias, since these households would have sent their children to school anyway. Or maybe only schools with particularly ambitious head masters voluntarily try out new pedagogical technologies. A positive correlation between the new technology and test outcomes may then be biased upwards, as schools with ambitious head masters may perform better on tests also for many other reasons than just this particular teaching tool. In these cases participation and technology adoption is referred to as "endogenous"; individual characteristics that may impact the outcome variable may also drive participation in the

program (selection), and both the outcome variable and the explanatory variable (technology) may be partly driven by a third factor that is hard to control for (head master ambition) causing omitted variable bias.

To get a clean estimate of the causal impact of the intervention, the evaluator needs strictly "exogenous" variation in the participation in the program, i.e. individuals should not get an opportunity to self-select into participation or not.[3] The solution to this problem suggested by the RCT methodology is to select a group of similar individuals/households/villages and then randomly pick a subset of these in which the intervention is introduced, leaving the rest as they are. To make sure that the randomization creates groups of treated and control that are similar, a baseline survey is undertaken before the intervention, and group averages for key variables are compared to rule out significant differences.[4] There will of course still exist large variation across subjects, but randomization guarantees that that variation is not systematically correlated to treatment status, and the focus in the evaluation is at the level of group averages. Once the baseline survey is done, the intervention is started, and after what is deemed as sufficient time for results to emerge (varies depending on

[3] Endogeneity is a very common problem in empirical work, and randomization is by no means the only way scholars try to deal with it. It is beyond the scope of this paper to discuss other approaches used, but using observational data and multivariate regression, methods include instrumentation, regression discontinuity, and propensity score matching. A short presentation of these approaches and how they compare with randomization is given in Duflo et al. (2007).

[4] Which these key variables are depends on the intervention at hand, the unit of analysis and the specific context, but typically captures socioeconomic aspects such as income, education, age, gender, etc. Sometimes scholars also go beyond just group averages and look at higher moments, such as the variance, to compare the two groups. Note that this can only be done for observables, there may of course still exist unobservable differences that correlate with treatment status. It is also common to stratify the units prior to the randomization to minimize the risk that the randomization still creates groups with significant differences across key confounding variables.

type of intervention and outcome variable to be studied) an end-line survey covering all subjects in both the treatment and control groups is completed. To estimate the impact, results from the end-line are compared across the two groups. In the cleanest cases, impact is derived simply from comparing average outcomes. Many times, though, multivariate regression including information from both the baseline and the end-line is used. This makes it possible to control for bias from observable confounding factors to the extent that the randomization has not succeeded completely in eliminating this. Using the baseline also makes it possible to single out the change over time in the treatment sample relative to the same change over time in the control sample, so called difference-in-difference results. In evaluations without an explicit control group, comparisons before and after an intervention may capture general trends in society that have nothing to do with the specific intervention. By having a control group and attributing the impact of the intervention to the difference in the change over time across the two groups, the effects of such general trends are controlled for.

The discussion so far has assumed that random allocation of actual treatment is possible. In some cases this is not the case, or at least not desirable, raising problems of selection bias. For instance, individuals cannot be coerced into taking loans or joining savings programs, making it difficult to randomize interventions such as micro credits and savings groups.[5] What can be done in such cases is an intention-to-treat (ITT) analysis, in which the intervention is offered to a

[5] There are innovative ways to come around this, for instance randomizing access to marginal applicants using credit scores as in Karlan and Zinman (2011).

random set of subjects (and not to another random set of subjects), and the analysis is based on the full samples even though it is known that not all subjects offered the intervention actually take it up. It is common in these cases that randomization takes place at a more aggregated level, such as the village or school, and classification as treated follow with that rather than individual take up of whatever is offered. To make it more concrete; to evaluate the impact of so called Self-Help groups in rural Rajasthan, Desai and Joshi (2014) looks at a random intervention of an NGO organizing such groups in 32 treatment villages, having 48 similar villages as a control group. However, membership within villages is of course voluntary, and it requires a reasonably steady flow of incomes, so there may be selection effects in who joins and who doesn't. Just comparing group members with non-members is thus not viable. Instead they compare women in treatment villages, group members as well as non-members, with women in control villages, giving an estimate of the intention to treat, i.e. the effect of having the option to take part in an organized Self-Help group. These effects may be more conservative, as the treatment group now includes subjects who have not been directly exposed to the intervention, but it can also capture possible externalities, good or bad, on others from the intervention.

The methodology has had immense influence on the field of development economics, and increasingly so on development policy. This influence comes from the premise that it can isolate an internally valid causal impact on the actual subjects from a range of interventions. From an academic perspective, much of applied empirical work relies on observational data, and problems of

endogeneity and selection are very common. Much of methodological development has been devoted to ways of dealing with this in a multivariate regression framework, and in every paper much effort goes towards convincing the reader that results are reliable given the inherit challenges. It is therefore not surprising that a methodology that offers a theoretically very simple solution to the problem, and promises clean and easy to understand results, gains a lot of traction.

It should also be an attractive tool from a policy perspective. Foreign aid is continuously questioned by voters and interest groups, and there is a pressure to show results and make sure that money is not wasted or stolen. In many cases an inability to point to results is justified by the notion that measuring actual impact of foreign aid is very complicated and many of the benefits are very long term, institutional in character and hard to quantify. There is no doubt some truth to that, however, impact evaluation can deliver numbers on at least some of the effects that aid financed activities yield, numbers policy makers can use to rationalize the size of aid budgets, and how that budget is spent.

Examples of applications

A full survey of papers having used RCTs in a development setting goes far beyond the reach of this paper. Below follows a few selective results across three areas where they have been influential; education, finance and corruption. The selection is not necessarily meant to be representative, and many important contributions are of course neglected. The purpose behind the selection was rather to illustrate through a couple of cases how broadly the methodology can be

applied, and to show its versatility and potential for actual policy influence.[6]

Education

That education is critical for increased labor productivity, and thereby economic growth and poverty alleviation, is disputed by few. There is less consensus though on how to best raise human capital, the relative role of quality versus quantity and supply versus demand side constraints. An influential approach to deal with demand side constraints are so called conditional cash transfer programs. The first such program was PROGRESA, in Mexico, offering poor families' cash transfers as long as the family seeks preventive health care and their children regularly attend school. This model was highly influential and spread like a firebrand across Latin America, and then beyond.[7] A crucial component of the political influence in this case was the ability to offer evidence that the program could actually work. This was particularly important in a country like Mexico in which programs are highly political and tend to be terminated when a new government takes office. From the perspective of the program developers, the Ministry of Finance, a less risky approach would probably have been to use methods more easily manipulated and with grounds for different interpretations, but instead they went for credibility and hard evidence. This was done through a pilot RCT

[6] For a good and easily accessible overview of the experimental work in international development see the book by Banerjee and Duflo (2012).
[7] How crucial the conditionality really is for the results have been questioned lately, as unconditional cash transfers have shown to have a rather similar effect elsewhere (Baird et al. 2009, and Benhassine et al., 2010). This suggests that what may have been driving the previous results were largely an income effect, rather than the effect of the conditionality.

implemented in a subset of villages, which showed that school enrolment increased from 67 % to 75 % for girls, and from 73 % to 77% for boys (Schultz, 2004). The successful pilot not only created political momentum to scale up the program to the rest of the country, but also gave it enough credibility to survive government turnover.

Another debate in education and development concerns how to raise school attendance in primary and secondary education (child absenteeism varies between 14 and 50 % in a set of surveys around the world reported in Banerjee and Duflo, 2012). Many obvious alternatives exist, and have been tried, such as reducing the direct costs of school attendance (tuition fees, or free provision of textbooks, uniforms or food) or improving the quality of education through teacher training or parent accountability mechanisms. One perhaps less obvious way to reduce absenteeism is through deworming.[8] A particular challenge for evaluation purposes in this case is that a significant part of the benefits of deworming are the positive externalities on other not directly treated children as their risk of getting sick also decline. This may suggest that an RCT is hard to undertake, but it turns out that it depends on how it is designed, and that there is quite a bit of versatility. In Miguel and Kremer (2004), the authors evaluate a program in rural Kenya in which a Dutch NGO together with the local government rolled out a free deworming program to 75 schools in three phases. The random assignment and gradual rollout (in each of three stages, 25 schools had mass-treatment

[8] The impact of deworming on cognitive skills and tests results had been studied before, but not the effect on attendance (Dickson et al. 2000).

reaching around 80 % of the children) made it possible to use an RCT even though full rollout was already planned from the beginning. Without any externalities, estimating the causal impact of deworming would have been relatively straightforward; compare average absenteeism among those receiving the treatment with those who didn't. With externalities only within schools, an intention-to-treat analysis could be made between treated and non-treated schools, which is also quite straightforward though it may generate somewhat conservative results. However, in this case externalities were likely even across schools, as neighbors, and sometimes even siblings, often went to different schools. This would mean that also the control group would benefit from the treatment, causing a likely underestimation of the beneficial effect of deworming.

To estimate and correct for this bias the authors used the natural variation from the randomization in geographical proximity of control schools to treated schools. Roughly speaking, the variation in outcomes across control schools that was systematically correlated to proximity to treatment schools was used as a measure of these externalities and used to correct the estimates of the impact. Taking these factors into consideration, the authors find that deworming increased school participation with at least 7 %, reducing absenteeism by one quarter.[9] They also used their results to calculate the cost efficiency of deworming relative to alternative interventions studied in previous papers. They calculate the deworming cost necessary to achieve an additional year of schooling to be $US 3.50, while the next

[9] Deworming also showed to have many other direct health benefits, but, on the other hand, the authors did not find statistically significant differences in test scores.

cheapest intervention (which paid for school uniforms in particular) cost $US 99.00. This paper highlights the versatility of the method also with complicated interventions and the ability to use multiple RCTs pilots to test what interventions are most cost effective.

Finance

As many other areas of life, development policy has its fads, fashions and hypes. One tool for development that has gotten a lot of attention the last decade (and even a Nobel Prize) is microcredits; small loans, largely but not exclusively to women, often with group liability, and typically with the intention to give the benefactor an opportunity to start or continue a small entrepreneurial activity. A substantial amount of resources, some of it from affluent philanthropists from the west, have been invested (numbers suggest that there are between 150 and 200 million micro-borrowers around the world), and pamphlets and webpages with heart-warming success stories exist in abundance. However, a recent wave of suicides among defaulting clients in India, and the increased role of commercial for-profit banks charging high interest rates have highlighted the need to get more systematic evidence on the average impact of microcredits beyond the success stories. This is not an easy task, though, for at least two reasons.[10] First, people applying for microcredits are likely to be different from those who don't, so just comparing borrowers with non-borrowers

[10] Another complicating factor is an apparent lack of interest to be evaluated from most of these institutions. In a way they are already working if they are financially sustainable, so why the need to evaluate their businesses? And, even more problematic, given the very favorable public perception, correct or not, there is really not much of an upside to being evaluated, as the risk of a disappointing result is greater than the benefit of confirming what people already believe anyway.

will suffer from selection bias. This suggests the need for an RCT, but you cannot force credit on people and it is unethical to deny credit to otherwise creditworthy clients. Second, at this point most markets are quite saturated with microcredit institutions, so even if a randomized trial can be set up with clients of one bank, members of the control group may have no difficulty finding a loan from somewhere else.

Despite these challenges, some impact evaluations are starting to come out. In Banerjee et al. (2013) they evaluate an intervention starting in 2005 in Hyderabad, India, in which a micro credit institution (MFI) entered half (randomly selected) of 104 defined slums. Other MFIs were in principal free to enter other areas, but there was quite little of that initially, reducing contamination of the control group. Evaluating differences across slums (since actual credits are self-selected) 15-18 months later, they find some evidence that entrepreneurs in treated areas invest more in their businesses, and that there are higher expenditures on durable goods. On the other hand, they find no increase in the number of businesses started or the profit of the existing enterprises. Furthermore, they find no impact on conventional development indicators such as consumption expenditures, health, education or female empowerment (these loans targeted groups of women). These results are largely consistent with similar studies undertaken in Morocco (Crépon et al., 2011), Bosnia-Herzegovina (Augsburg et al., 2012), Mexico (Angelucci et al., 2012) and Mongolia (Attanasio et al. 2011). Does this mean that microcredits have no impact? No, but it is not the miracle solution sometimes argued, just one of potentially many tools of the trade. Considering the small size of the credits, the typically very strict

conditions on repayment, the two-edged sword of group lending, and the focus on zero defaults, these results are not really that surprising, just seemingly so given the hyperbole. And knowing this is of course very important for aid agencies having to make tough decisions on how to allocate their limited resources. This work thus once again offers some important policy implications, and the methodology was critical for finding, and building credibility behind, these somewhat controversial results.

Corruption

Until recently corruption was regarded as something observable but not quantifiable, stemming attempts at empirically estimate causes, consequences and the effectiveness of anti-corruption policies. Lately both macro (cross-country) and micro (firm or household) level data have become available, creating a large and growing literature. In the literature on anti-corruption policies, two different directions can be identified, one focusing on accountability from above, the other on accountability from below. Substantial amount of money has been spent on community development programs with the ambition to help poor communities strengthen their ability to hold local politicians and providers accountable for poor public service delivery generally, and corruption specifically.[11] In many cases, however, little is known of how effective they are, and even less about the effectiveness of this kind of accountability from below relative to that from above.

[11] For instance, the Indian government together with the World Bank are spending 8 billion US$ to organize 150 million households in Northern India into so called self-help groups, one of the objectives being that this will help poor communities hold service providers accountable to their actions.

In Olken (2007), the author looks at a rural roads project across a set of 608 villages in Indonesia. Infrastructure programs are known to be prone to corruption, and Indonesia is a high corruption environment. Top-down approaches are therefore vulnerable to capture, and hopes may instead rest on grassroots initiatives. To test the impact of top-down monitoring, a group of villages where randomly selected to have the probability of an external government audit increase from 4 to 100 % (and this was publicly announced). To test for community participation, two experiments were conducted, both aiming to increase active participation at so called "accountability meetings" at which project officials account for how they have spent the money. The first treatment involved hundreds of invitations to these meetings being distributed throughout the village. In the second treatment, an anonymous comment form was distributed along with the invitations. Forms were collected before the meetings in sealed drop-boxes, and the results were summarized in public at the meetings. Both interventions raised grass-roots participation levels in meetings substantially. To measure corruption, Olken calculated missing expenditures by comparing officially reported costs for material and wages with estimates rendered from road core samples and village surveys. Somewhat surprisingly, Olken found that the top-down approach (even though findings of corruption rarely led to any legal action) led to a significant reduction in missing expenditures by 8 %, while the participatory approach had a very small and insignificant overall effect. How can this be? Disaggregating the results, he found that the participatory approach had a significant effect on missing wage payments, but none whatsoever on material costs, the largest

share of total expenditures. He interprets this as a free rider problem; individuals stood to gain personally from raising complaints when they had not been paid as much as reported, whereas the benefits of reducing leakage in material expenditures was shared between all village members. This highlighted a critical component of programs aiming at raising accountability from below; encourage trust and cooperation to reduce the inherit challenges to collective action. The versatility of the methodology made it possible to test both types of approaches within one unified context, and highlighted a key challenge a popular type of program is likely to face also elsewhere.

Strengths and weaknesses

Scientific progress to a great extent builds on questioning current research. So when a new approach, school of thought or methodology becomes influential, it also comes under scrutiny. This is true also for the application of RCTs to international development. In this case the scrutiny probably has found extra fuel in the perception that some of the strongest proponents of the methodology have been perceived as touting the benefits of the methodology at any cost while being dismissive of any alternative approaches. For instance, it has been argued that aid money should exclusively be channeled to projects that have been shown to have an effect through "hard evidence" (read RCT), since results from alternative evaluation methods used in the social sciences are dismissed as lacking internal validity (Banerjee 2007). A common complaint is that the so called "randomistas" are monopolizing the field, dismissing all alternative methods as useless, and ignoring many pieces of evidence and information that are policy relevant because they do not fit the methodology. On the other hand, proponents of the methodology argue that policy decisions must be evidence based to avoid arbitrary decisions (think about the Jeffrey Sachs v. Bill Easterly debate, and in particular the recent discussion around the Millennium Villages Project), and if so, there is no reason not to use the best methods to acquire that evidence. Below follows a discussion of the strengths and weaknesses of the methodology based off some of the critique that has been brought up, but also the response to that critique. The purpose is to offer a better understanding of limitations and strengths of the methodology, which

is crucial when thinking about when and where to use it for policy purposes.

For policy analysis it is useful to make a distinction between two different purposes of impact evaluation. The first purpose is close to that of the academic literature, using RCTs as part of the experimental approach to learning about general mechanisms and interventions that foster development. In this case the evaluation is very explicitly meant to speak beyond the immediate context of the intervention, and should offer some guidance in terms of what types of interventions that are effective and efficient. This knowledge is important for an aid agency's strategic work when planning future interventions and areas of focus, and should have an agency-wide interest. The second purpose is more immediately related to understanding the impact of a particular intervention. This may be relevant when discussing whether to start a new project/program within a specific setting, but also when considering whether to terminate or continue an already existing project/program. As the purpose here is more narrow, some of the concerns and critique brought up below may be less (or in some cases more) relevant than for the more general purpose. I will therefore discuss also to what extent the limitations brought up below apply depending on the underlying purpose of the evaluations, which I will refer to as long run versus short run purposes for lack of better words.

Internal validity

The main advantage of randomization is that it helps with internal validity, i.e. a scholar can feel quite certain when making causal statements based on the evaluation results. This is because

randomizing the treatment across similar subjects reduces systematic problems of selection and bias from confounding variables, so the causal effect of the treatment can be identified. This is not the only way to theoretically identify a causal impact, but it generally requires fewer specific assumptions about individual behavior, the specific environment, etc., and is therefore more robust to our lack of knowledge of how reality exactly works.[12] However, proponents of the approach often seem to go one step further and claim that the methodology guarantees internal validity. As pointed out in e. g. Deaton 2009, Ravallion 2009 and Rodrik 2008, identification still relies on certain assumptions that may be violated. In many cases spill-overs into the control group cannot be ruled out (or accounted for as in Miguel and Kremer 2004 discussed above), or members of the control group may change their behavior when they realize they have not been selected for what seems like an attractive opportunity (there is no placebo in social experiments).[13] Another concern is so called substitution bias. This occurs when the control group may have access to a close substitute to the treatment. Think for instance of the case

[12] This is the common view, but there are contrary views on this as well, e.g. Heckman and Smith, 1995.

[13] Those selected for treatment can also change their behavior in unexpected ways that contribute to the outcome beyond the actual intervention. These so called Hawthorne effects are commonly discussed in psychological, anthropological, and even behavioral economics studies where selection is not double-blind and the presence of observers may alter the behavior of the observed. A recent study by Bulte et al. (2012), compared outcomes across three groups, conventional control and treatment groups and a double blind group not knowing whether they had received the improved seeds that were studied or not. They found that the whole positive effect on yields comparing the treatment group with the control group also showed up when comparing the double-blind group with the control group, despite half of the double-blind group having the conventional seed. They attribute this to a behavioral response; both the treatment group and the double-blind group planted seeds with greater distance relative to the control group, and this, rather than the quality of the seed, led to higher yields (but possibly also less of other produce as more space was taken up). Without the double-blind group, the positive effect would have been attributed to the seed variety, when in reality it came from a behavioral response that only the combination of all three groups could pick up.

with micro-finance. Even if access to a loan from a particular micro-credit institution (MCI) could be randomized, if the market also has other available MCIs, then many of those denied can get a loan from someone else. Members of the control group thus get access to a close substitute to what is available to the treatment group, so the average difference across the groups cannot be seen as the average treatment effect (which is supposed to measure the difference in outcomes for the same individual getting a loan versus not getting a loan).

It is important to understand that alternative methods of evaluation also suffer from these problems, so this is not so much a critique based on the inferiority of this method relative to alternatives, as it is a critique against those arguing that this method solves all problems. It is a critique to take seriously, but to use it as an argument for choosing an alternative approach to evaluation, an explicit argument needs to be made on how the alternative method outperforms RCT in this respect. The critique does thus not necessarily suggest that the methodology is inferior to alternative methodologies in this dimension, but rather that some caution is warranted also when interpreting results derived under randomization.

The micro-credit case above is also a good example to illustrate the difference between the short term and long term objectives of impact evaluation. The long term objective would be to learn more about the general value of getting access to credits. In this case the evaluators want to make sure the control group has no access to credits or close substitutes, and they need to think seriously about external validity. An aid agency may certainly be interested in this general question to guide future policy initiatives, and get a sense of the value of focusing

resources on micro credits rather than something else. However, it may also have an interest in an impact evaluation simply for the short term objective of finding out whether this particular project, given market saturation, demand for loans, etc., is working. In this case you want to estimate impact including the substitution bias as this is a characteristic of the environment in which this project is operating that most likely has important implications for the value added of this particular project.

External validity

A well-known problem with RCT's is that of external validity, i.e. the extent to which results can be generalized.[14] Most experiments are undertaken within a confined environment so the effect may be contingent on factors common among treated and controls, or factors specific to the way the intervention was implemented. This is something that randomized experiments by design have a hard time picking up. After all, the idea is to find treated and controls that are as similar as possible with respect to all factors, other than the treatment, that may influence the outcome. But, it means that it is hard to know if what seemed to work in Tanzania, would also work in Laos.

Also note that concerns with external validity do not just reflect differences across participants and the economic, social, cultural,

[14] Even proponents of RCTs acknowledge external validity as a concern, but still papers using RCTs typically spend very little effort on discussing this challenge, and what it means for the policy implications of the study at hand. As pointed out in Rodrik (2009), in studies using observational data authors typically spend a significant amount of effort and reasoning to motivate their specification to reduce concerns for internal validity, perhaps the key challenge for that methodology. In RCT studies, though, the key challenge of external validity is often not even mentioned.

geographic and institutional context in which they operate, but also with regards to the identity of the organization implementing the program, and the political situation in the partner country. This also has implications for what to expect in case the project is scaled up. Is it reasonable to believe that a program scaled up to the national level, and with authority transferred from a NGO to say regional governments, will be implemented in a similar way as the initial small pilot directly overseen by an evaluation team?[15] There are also effects that may not translate when scaled up. A small scale experiment estimating the returns to an additional year of schooling (derived for instance from deworming) takes wages as given, but if this would be achieved at a national level, the relative return to skills and education would of course also change (Acemoglu, 2010). Scaling up typically also brings in an additional political economy dimension, where distributional conflict may arise, corruption in program implementation can become a concern, and the allocation and use of the program can be hijacked for political purposes.

To deal with these limitations replication studies in different environments help. However, as pointed out in Banerjee and Duflo (2009), a challenge to replication studies is that they yield less academic credits and may seem less attractive in the eyes of funding agencies looking to fund high-visibility projects. This suggests that donor money may be needed to create knowledge banks of impact

[15] A recent paper, Bold et al. (2013), show exactly this point by randomizing the implementation of an education program in Kenya across an international NGO and the Kenyan government. Relative to a control group with no intervention, the NGO implemented program showed a significant improvement in test scores, whereas the government implemented program showed no significant improvements. The authors attribute the difference in results to a combination of political economy factors and lacking implementation and monitoring in the government run program.

evaluations and that it is not possible to rely solely on scholars to drive the process. It should also be noted that single replication studies don't invalidate all arguments for concern. Variation in impact across environments is typically quite high, so meta-studies have to rely on a relatively large number of replication studies to get statistically reliable results on average effects. There is also recent research, comparing estimated impact across different environments using RCTs, as well as within the same environment using RCTs and econometric methods using observational data. Pritchett and Sandefur (2014) find that using a different method but within the same environment generally come closer to the results from the RCT in that same environment (taken as the true estimate of environment specific impact) than RCT studies on the same intervention in a different environment. Their approach is not without fault and results need to be confirmed by more studies, but the authors interpret this as a serious disclaimer against the argument that evidence based policy decisions should only rely on internally valid methods, i.e. RCTs. Many times, their argument goes, alternative methods in the same context serve as a better guide, and development agencies must trade off internal and external validity if ranking evidence going into their decisions.

It should once again be emphasized that external validity is an important concern for learning about the effectiveness of types of interventions, and if donors are considering scaling up current interventions. However, if an aid agency is primarily concerned with the short term objective of impact evaluation, the effectiveness of a particular project/program, it may be less crucial. If similar interventions are planned elsewhere, and the RCT is used to evaluate

whether this would be a good idea, then external validity is crucial. If the question is whether the current project should be continued or terminated, then of course external validity is less of a concern.

Selection

Obviously not all types of interventions can be randomized, or at least some would be prohibitively costly or make no sense for other reasons. In principle this should not be a problem; use RCTs whenever they can and should be used and use the best alternative method when they cannot. The concern, however, is that the attitude that only "hard" evidence counts will push development policy and resources towards interventions that can be randomized even if they may not be the interventions we have reasons, ex ante, to believe are most important for development. Large infrastructure projects, for instance, may have large social rates of return, but they are typically quite costly, though not impossible, to randomize.

Another selection problem comes from the need to find facilitators to work with. A recent study by Brigham et al. (2013) sent out an invitation to microfinance institutions to form a partnership to evaluate their programs. With the invitation followed a survey of previous findings in two versions. Half of the institutions were randomly selected to get a survey indicating positive impact of microfinance, half got a survey indicating no effect. The number of responses in the first group was twice that in the second group. This suggests that self-selection of organizations may determine what programs are evaluated, and a confirmation bias may emerge if

organizations with a more positive prior of their effectiveness are more willing to be evaluated.

Selection goes beyond just types of interventions and organizations to cooperate with, though. RCTs are developed to measure the average treatment effect of a well-defined intervention with observable and quantifiable outcomes.[16] Many times this may not be the only outcome of interest, though. Policy makers may be interested in diverse effects across different groups (something that in principle is possible with RCTs but often run into problems of statistical power given the limited samples), details on the implementation and collaboration with the facilitator, learning within the partner organization, and the long run sustainability of the institutional context in which the intervention takes place (and the contribution of the intervention to the strength of that institutional context). Most impact evaluations generally also have a fairly short time horizon in order to avoid contamination of the control group. This implies first of all a selection of short term results over effects that may take longer to evolve, but that are equally important. Second, it may also be that short run results are unsustainable once the original effort fades out, something a conventional impact evaluation may not pick up.

The methodology may thus also dictate a selection in terms of what outcomes to focus on, in particular towards more short run and

[16] The literature makes a distinction between the average treatment effect on the treated, and the average intent-to-treat effect. The first measure counts only those who directly get the treatment as part of the treatment group, while the other counts all who were given the opportunity to take up the treatment as part of the treatment group. Think of a savings group intervention randomized at the village level, but where only some dwellers chose to become part of a savings group. According to the first definition only those becoming members of a savings group are part of the treatment group, while in the second case all village members in treatment villages are part of the treatment group.

easily verifiable and quantifiable outcomes over more long run and harder to quantify effects. Proponents of the approach would say that this is just a good thing. The allocation of resources should be guided by hard evidence on what works and what doesn't, and why spend resources trying to understand effects we cannot measure with any precision anyway. Hard evidence fosters discipline, whereas current allocation decisions often are guided by sloppy thinking, gut feelings and political considerations, all made possible by the argument that we cannot know for sure what works and what doesn't anyway. Skeptics, on the other hand, may argue that development is a long run process, not something that happens one experiment at a time, and that learning and institutional development may be at the core of what aid is trying to achieve (even when targeted towards a very specific activity). This may be a particular concern for policy purposes if it reinforces a political bias in favor of easily verifiable results that can be sold to a population suffering from aid fatigue (more on this below).

Ethics

A common concern among practitioners regards the ethics of the approach. After all, members of the control group need to be denied access to the treatment, as well as close substitutes, for the time it takes to evaluate the intervention, which sometimes extend to a couple of years. If the treatment was known to be effective and efficient, and resources were no constraints, then the ethical thing to do would clearly be to roll out the program to everybody. This is not typically the case, though. First of all, the evaluation is done because of the fact that impact, and efficiency relative to alternative interventions, is

unknown. Ethical concerns can emerge though if preliminary midterm evaluations suggest a very strong positive effect (or a negative effect for that matter), in which case the ambition to finish the full evaluation before rolling out the whole program (or cancel the trial) may raise an ethical dilemma.

Second, there are typically financial and other constraints suggesting that there is only a sample of potential beneficiaries being targeted initially anyway. Many times RCTs take advantage of interventions that are already from the start planned to be rolled out sequentially due to financial and/or human constraints (such as the case of deworming discussed above). In this case the ethical concern has more to do with how that selection process is designed than with the fact that not all are covered right away. In particular, is it ethic to randomly assign treatment when practitioners believe they have local knowledge specific enough to target groups for which the intervention is particularly important or is likely to have a particularly large impact? RCTs are often stratified across one or a few observable dimensions, such as income or family size, but NGOs and other facilitators may have access to more local and hard to quantify information that cannot be used for that purpose (Ravallion, 2009). With limited resources available, doesn't the ability to have a large impact through selection of beneficiaries trump the potential learning benefits of average effects through an RCT? There is no obvious answer to that question and the relative merits of either approach is case specific and depends on factors such as the ability of the facilitator to correctly target the intervention and the future ambition to scale up the intervention (in which case learning is relatively more important). On the other hand,

it is a well-known problem that social interventions sometimes get targeted towards groups based on their political connections/ importance or economic strength, rather than based on their needs or anticipated effectiveness. In such an environment, randomization should be regarded as the ethical alternative, breaking the link between connections or wealth and publicly financed services.

Resource requirements

A final concern brought up here are the human and financial resources necessary to do RCTs. Properly implemented most RCTs require that the evaluation team is involved already from the very start of the intervention to do the randomization and the baseline survey. It is also critical that the baseline includes all relevant questions as this cannot be corrected for ex post, and that all partners involved in the intervention do their part. Incentives may not always align, and it is important to guard the integrity and interest of participating NGOs and other local participants. RCTs can also be quite costly. Typically two rounds of surveys are necessary (though a baseline is not always necessary), preferably reaching out to a relatively large sample to guarantee statistical power of the evaluation. In addition, it may be necessary with a pre-pilot to test the survey instrument or other aspects of the study.[17]

[17] It is difficult to put an exact number on the cost of an RCT as it depends among other things on the country studied, the necessary sample sizes, and the need for a pilot. RCTs also often require smaller sample sizes compared to quasi-experimental evaluation methods, as the stronger internal validity helps with power also at smaller samples. The cost of the RCT should also be contrasted to the risk, and thereby cost, of making decisions based on less reliable information (http://www.poverty-action.org/about/faqs).

All types of evaluations will of course require resources. It has been argued, though, that for the reasons mentioned above, RCTs may be particularly costly and require more planning than most alternatives. This may be particularly true if the evaluation primarily has a short term motive. If in the end all that is learned is about the impact in a very specific context, then that knowledge may not seem enough to motivate the full cost of the approach. This cost must be put in perspective, though, to the cost of misguided projects and programs. It is difficult to put a number on the costs of continued spending on projects with no impact. Nevertheless, given the rather modest difference in evaluation costs, and the quite substantial amounts of money invested in projects and programs, RCTs do not need to increase the chances of terminating ineffective interventions with much in order for them to be cost effective.

Still, one way to make RCTs more cost effective would be to bridge them closer to theory, and structural models (Acemoglu 2010, Heckman and Smith 1995). These models, if correctly specified, can be used to simulate the effects of different interventions in different environments, a way to get at external validity without the need to extend the full cost of additional RCTs. The original RCT is then used to produce structural parameters that are fed into the theoretical model, and the sensitivity of the results to different environments can be estimated. This was largely how RCTs were originally used in labor and public economics, but it does require model assumptions, something that fell out of fashion.

When to consider an RCT?

So, given the discussion above, under what circumstances should an aid agency or donor country government consider an RCT evaluation? It should first be emphasized that commissioning your own RCTs is not the only way to learn. Taking advantage of the large and growing number of existing RCTs is equally important. The first question to ask then is if similar interventions have been undertaken elsewhere and if there has been credible impact evaluations done. If so, then the next question is what can be learned from these earlier studies given external validity concerns. Is the context of the intervention considered very specific, or at least substantially different from those where previous impact evaluations have been undertaken? In the best case scenario several evaluations are available which helps greatly with analyzing the generalizability of results. If not, a careful assessment needs to be done, based on differences in the socio-economic, institutional, geographic and cultural context, and how crucial these differences are likely to be for the results. As suggested in the paper by Pritchett and Sandefur (2014) discussed above, evaluations using other quantitative methods in the same environment can also be useful.

Nevertheless, there are situations when commissioning a new RCT seems well motivated. Below is a tentative list of conditions under which this is particularly true.

1. Few credible previous evaluations are available so the value added of new information is high.

2. The intervention can be randomized, and randomization does not incur unreasonable additional costs.
3. The intervention is randomized from start or is not yet fully implemented, and the time frame to decision allows for the time required for impact to be observable and measurable.
4. The intended impact(s) of the intervention can be observed, measured and quantified in a way compatible with the methodology.
5. The intervention is planned to be scaled up (terminated) if the randomized trial shows satisfactory (un-satisfactory) impact.

A few of these points are obvious, but some are worth some more explaining and discussion. Many aid financed projects/programs cannot be randomized, at least not in their entirety. These programs are sometimes referred to as complex and often fall under headings such as democracy support, institutional development and governance. It should be noted, though, that there are often elements of interventions also within these fields that can be randomized. There is an abundance of papers using RCTs looking at for instance the effect of information campaigns, village group organizations, and gender quotas on different aspects of governance and political accountability and transparency (e.g. Björkman-Nyqvist and Svensson, 2009, Desai and Joshi, 2014, Duflo and Chattopadhyay, 2004). That is, even with complex programs, randomized trials can often be used to inform us

about how the design of the intervention influences impact.[18] Nevertheless, DFID (2012) discusses alternative methods for impact evaluation that may be more useful in such complex contexts. These different approaches have different pros and cons that are related to internal versus external validity among other things. Many of these methods are yet to be tested out in a development context, but they offer an interesting complement to RCTs in complex settings.[19]

The planning required and time frame needed to perform RCTs can be a concern for policy purposes. Ideally an intervention should be randomized from start, but even if it is not, it is often possible to use so called encouragement design to get random variation in take up after the fact as long as the intervention is not fully implemented and has reached all of the intended beneficiaries. For instance, a targeted marketing campaign for a microcredit bank can include a lottery ticket that gives recipients a 50 % chance to open a free savings account with some small amount already deposited if they apply for a microcredit loan. As long as the opportunity is randomly assigned, and the opportunity is attractive enough to really spur a substantial increase in applications in the treatment group (those winning the lottery) relative to the control group (those losing the lottery), then the

[18] A neat example within politics is Banerjee et al. 2011. They randomized an information campaign across slums in Delhi that provided subjects with information about the performance of the current local political incumbent, and the qualifications of the incumbent and two main opponents in an upcoming election. They found that treatment slums had higher voter turnout, less vote buying, and a higher vote share for better performing incumbents. This suggests that aid financed democracy support to reduce information asymmetries can be a very useful tool to increase accountability from below, and the RCT was essential to acquire that knowledge and make it credible.

[19] The question of alternative approaches to conduct impact evaluation and assure causal inference with complex interventions is very important and deserves more attention than I have scope to offer here. See DFID (2012) for a more in depth discussion of these alternatives.

encouragement can be used as an instrument when estimating a causal effect of micro credits. Hence, initial randomization is to prefer, but it is not a necessary condition to do an RCT. It should also be emphasized that policy concerns are often a reflection of a planning horizon that is too short and too focused on process and output rather than impact. As discussed more below, impact evaluation can enforce more discipline into activities already at the planning stages as it requires a very clear definition of what the objectives of an intervention are, what impact is expected, and under what circumstances an intervention should be terminated, or not scaled up.[20]

That the intended objectives can be observed and measured in a way compatible with the methodology points to the fact that RCTs typically only offer answers to part of the questions an aid agency is interested in. This doesn't necessarily mean that an RCT is not warranted, but it suggests that it often needs to be complemented by other methodologies that can better get at these other questions. Using complementary methods, so called mixed methods, drawing on the strength of each, is becoming more and more common (e.g. Bamberger et al. 2010). Mixing impact evaluation with focus group interviews, field experiments and analysis of Management Information System (MIS) data can help the evaluator getting answers to a broader set of questions (related to both impact and process) and also better

[20] RCTs may also reduce the dimensionality problem of the bureaucratic tasks in agencies overloaded by New Public Management routines. My impression is that the idea that final impact cannot be measured with credibility is partly to blame for an excessive number of indicators of intermediate outcomes, output, and fiduciary management that has to be tracked, measured and reported.

understand why impact, or lack thereof, is observed.[21] Qualitative methods (methods of inquiry that rely more on open ended questions, and less structured approaches) can also help with external validity by identifying contextual components critical for the results. It is thus important to not necessarily see the methods as substitutes, but rather as complements.[22]

Finally, as discussed above in terms of the short term and long term objectives of impact evaluation, an RCT can be motivated even in the absence of external validity and a bigger picture if the value added of a more precise estimate of the impact of a particular project/program exceeds potential additional costs or delays (for instance if a decision on whether to terminate a project/program is pending). However, the value of the RCT is of course higher if it has ramifications beyond the specific context considered. An ideal case is to evaluate a pilot of a larger planned national intervention to be rolled out more broadly (think of the PROGRESA program in Mexico), keeping in mind the results in Bold et al. (2013) on pitfalls when scaling up. An alternative is to evaluate a project/program in one country if considering similar interventions in another country. Once again there are pitfalls in making inference across different environments, so a careful analysis of external validity is necessary.

So, are there no risks of emphasizing the need for RCTs for project and program evaluation? There are, if it leads to an excessive reliance on this as an exclusive method of evaluation, if the methodology

[21] Results from RCTs are sometimes accused of being black box, that there is no clear theory attached that can explain why the results observed emerge (e.g. Ravallion 2009).
[22] The merits of using mixed methods deserve more attention but goes beyond the scope of this paper.

comes to determine the types of interventions that are tried, and if it reinforces existing political biases. That the methodology cannot give answers to all relevant questions has already been emphasized. The second concern relates to the argument sometimes made that aid budgets should only be allocated based on hard evidence of what works and what doesn't based on RCTs. If so, then there are many types of interventions that would never even be tried simply because they cannot practically be randomized, and therefore we cannot have any solid RCT based evidence of their impact. An aid agency's program portfolio needs to be determined by a broader set of criteria of need and ex ante potential impact (building on theory and alternative methods of evaluation), and of the course the partner country priorities.

The political bias concern is explained more in detail in Olofsgård (2012). The argument builds on the assumption that in a first best setting in which the principal in charge of aid allocation is only governed by development impact then a better methodology for short term impact evaluation can only be a good thing. But, in a second best world where the principal, for instance because of political visibility, already has a bias in favor of what is short term quantifiable, then an improved methodology to study just that may cause a bias against activities with more long run institutional effects that may be at least as important for development. This argument is related to the discussion above that RCTs cannot be used to answer all questions, but it also highlights that this is not necessarily a concern unless there is already a bias in the allocation decision. Such bias cannot be ruled

out, though, fed by the need to motivate generous aid budgets by showing simple and tangible results to media and voters.

RCTs and Swedish Aid Policy

In this section I will firstly give my impression of what the current situation looks like when it comes to RCTs and high quality impact evaluation in Swedish aid. I will then motivate why I think it would be valuable to bring in RCTs more into the evaluation toolbox. Finally, I will offer a brief discussion of how that could possibly be done, and what it would require.

The Current Situation

The Swedish management of aid evaluation has undergone major changes lately. One of the most dramatic events is of course the decision to terminate SADEV that came in the fall of 2012, after a critical report from Statskontoret (2012). But, aid evaluation of course also takes place within Sida, an organization that has been under pressure due to reorganizations and budget problems.[23] Most importantly, the division responsible for evaluation at Sida, UTV, has been restructured twice during the last decade, from a quite independent department reporting directly to the board of directors before SADEV was created, to a sub-department partly tasked to support operations and reporting to the General Director.[24] This is by no means necessarily a bad thing, but it is easy to get the impression that even though the importance of showing results have been much

[23] There are several public agencies that occasionally evaluate expenses that fall under the aid budget, but as pointed out in Statskontoret (2012), these evaluations are typically of marginal overall importance or of an ad hoc character.

[24] A current change taking place is the appointment of a new chief economist, starting work on March 1, 2014. This may potentially influence how evaluations are conducted and organized in the future.

emphasized by politicians, in practice the task to make sure that aid evaluations are of good quality has not been given political and administrative priority. The question is if the institutions and individuals tasked with the responsibility have been given the resources, incentives and environment necessary to incorporate new methods and thinking.

In a global perspective, RCTs have during the last decade gone from being primarily an academic preoccupation to an important policy tool for leading actors in the international development community. Both bilateral and multilateral donor organizations (including Sida) have invested substantial amounts into organizations undertaking and promoting best practice impact evaluation such as 3ie, the Abdul Latif Jameel Poverty Action Lab (J-PAL) at MIT, and Innovations for Poverty Action (IPA) at Yale University. At the World Bank there are also trust funds explicitly devoted to best practice impact evaluation, such as the recently closed Spanish Trust Fund for Impact Evaluation, and the multi-donor Strategic Impact Evaluation Fund. The British aid agency, DFID, and the US counterpart, USAID, have been particularly active, and so has many private charities, such as the Bill and Melinda Gates foundation.

Surprisingly little of this is seen in the Swedish aid community. Statskontoret (2012) noted that SADEV had not undertaken a single impact evaluation of any kind, even less an RCT. A report from 2008 on evaluations at Sida UTV (Sida, 2008) made the following assessment of the methodology used in 34 evaluation reports under study (See Table 1). As can be seen, no quantitative evaluation methods had been used, and certainly no RCTs. This doesn't mean

that these evaluations didn't discuss impact, almost half of them did. The concern, though, is the quality of that analysis, and in particular the ability to credibly weed out causal impact from the interventions. Or, as stated in the report: "Impact analysis would in many cases require stronger designs to generate valid and reliable conclusions. We have to conclude that the selection of methods and sources of data collection were not adequate" (Sida, 2008, p. 74).

Table 1 Designs chosen in the evaluations

Design alternative	Number	Percentage
Randomized control group pre-test – post-test design	0	0
Non-randomized groups pre-test – post-test design	0	0
One group pre-test – post-test design	0	0
One group time series design	0	0
Judgmental sample, case study design	12	33
Narrative analysis	22	67

Source: Sida (2008), p. 49.

This report came out 6 years ago, but according to a more recent report (Transtec, 2014) little has changed with regards to the use of quantitative methods and the quality of impact evaluation. This report was commissioned to evaluate Sida's Framework Agreement for reviews, evaluations and advisory services, but as part of that evaluations are analysed and rated based on a number of criteria. The report also includes an explicit comparison with the findings in Sida, 2008, and identifies some limited improvement in a few areas but not when it comes to quantitative methods. As stated in the report: "The most prominent deficiency in this regard is the lack of applied methodologies for data collection and analysis. This is the basis for 'evidence based' evaluations and is critical when evaluating any level of

evaluation or for distinguishing causality within a theory of change or from input to outputs and outcomes" (Transtec, 2014, p. 59).

It is difficult to get a fully reliable sense of the extent to which Sida has commissioned RCTs. From what I have been told in informal interviews, there are projects that Sida have been part of financing that have had components evaluated using RCT, but these have not been commissioned by Sida itself.[25] Looking at Transtec (2014), none of the projects listed there have used an RCT as far as I can understand from the information given. I could identify 75 evaluations that had been rated in terms of whether there has been an accurate assessment of impact. Of these, impact assessment was deemed not applicable in 18 cases, and in 15 cases no impact assessment had been done at all (even though it was deemed applicable). The remaining 42 evaluations were rated between the lowest and highest possible scores, with an average assessment of 3.83, where 3 means "not quite adequate", and 4 means "minimally adequate". It is not obvious what to make out of this since there is a lot of subjective judgment involved, and it is theoretically possible than an RCT or other experimental approaches would have been a bad fit in all these cases. I do find the consistent weakness in quantitative methods, and total absence of any apparent attempts at doing an RCT, striking, though. It is also valid to wonder what it says about the choice of projects/programs that Sida gets involved in, if it is the case that an RCT was never the plausible method.

[25] An interesting academic example of an RCT done on Sida's work is Bengtsson and Engström (2013). This paper looks at the effect of formal modes of monitoring of NGOs contracted by Sida (otherwise relying on a trust based system). The authors, randomly assigning the formal monitoring system across contracting NGOs, find that monitoring increased outreach, reduced expenditures and reduced also financial irregularities within the NGOs. The study was made possible by the help and support of Sida, but it was financed by The Swedish Council for Working Life and Social Research.

It is also possible to learn something from Sida's own instructions to their staff. Sida's manual for evaluation of development interventions, "Looking Back, Moving Forward: Sida Evaluation Manual" discusses evaluation on more than 100 pages. It brings up both effectiveness and impact under the rubric "Evaluation Criteria". The language in the manual uses many of the concepts necessary for robust impact evaluation, but, across the 3 177 words under that rubric, never mentions randomization, trial or RCT.[26]

So, why has Sweden, a country traditionally priding itself as being a competent and unbiased donor, fallen behind many other donors in adopting new methods? There is no doubt that knowledge exists about these methods, and about the merits of these methods. However, it is easy to get the impression that despite the focus on results in Swedish aid the last few years, the ambition to incorporate this as a tool in the Sida toolbox just isn't there yet.[27] I can only speculate about why this is the case, but I can see that the incentives

[26] Sida's webpage, http://www.sida.se/English/About-us/How-we-operate/Sida-Evaluation/Manualer/, offers a link to what they refer to as a good example of an impact evaluation, a research paper by Björkman-Nyqvist and Svensson (2009). This evaluation is partly financed by Sida, but the project analyzed is financed by the World Bank. From what I can tell, no similar RCT impact evaluation has been commissioned by Sida for any of their own projects or programs.

[27] That knowledge exists, but expectations and ambitions fall short is keenly illustrated in the following quote (Sida 2007, p. 34): "The second main task is to decide, with as much certainty as required or possible, whether the changes that have occurred since the beginning of the intervention were caused by the intervention, or if they would have occurred anyway. Impact, in the strict sense, is the difference between the changes that have actually occurred and the changes that would have occurred without the intervention. The hypothetical state of affairs to which we compare real changes is known as the counterfactual. With the help of control groups that have not been exposed to the intervention it is sometimes possible to get a good idea of how the target group would have fared without the intervention. When the counterfactual cannot be estimated in this way – a common situation in development co-operation – statements about impact rest on weaker foundations. The intervention is often taken to be the cause of the identified changes if such a conclusion appears to be consistent with expert knowledge and there seems to be no better explanation around. Although less compelling than an explanation based on control group methodology in most cases, an argument of this type can be good enough for the purpose of the evaluation."

of individual desk officers to push for this are slim. In particular in a situation of turmoil and reorganization, the time horizon shrinks, and focus shifts towards the most immediate needs, and not towards new routines that require more planning, a longer time horizon, and possibly difficult discussions with partner country counterparts and implementing NGO's or consultancy firms. A change must thus start from the top, and be part of an organization-wide push towards improved methods of evaluation generally (with RCT's being part of that). This would of course also require that Sida gets support for this change from the government and the Ministry for Foreign Affairs. The ability of any government agency to work long term and structured in the end also depends on the hand they are given by their ministry. In this case this involves the objectives of aid as defined by the government, the resources made available to evaluate and monitor, and the understanding and patience with methods of evaluation that may take somewhat longer to yield answers.

The Benefits of More RCT in Swedish Aid

I would argue that there are at least three reasons why taking impact evaluation more seriously would benefit Sweden as a donor, which in turn would benefit also the intended beneficiaries. First, the obvious one, it is important for aid effectiveness to know what works and what doesn't and RCTs is an important tool for that purpose (though far from the only tool). It is important to get away from the view that impact evaluation is research, and not all that useful for policy

purposes.[28] Impact evaluation is very concrete, and can offer tangible results directly related to the impact of aid financed activities. Aid fatigue is a valid concern for those who believe that aid is an important tool to promote development, and RCTs yield tangible and concrete results that should be possible to communicate to a broader public.[29]

Second, Sweden's influence and reputation within the international donor community is likely to suffer if we are lacking the competence and/or ambition to incorporate what is considered as best practice impact evaluations into our toolbox. Sweden is a generous donor with high ambitions and a reputation for having "pure" motives (less motivated by strategic, commercial or old colonial ties than some other bilateral donors). The generosity, however, necessitates a certain responsibility towards tax payers as well as the intended beneficiaries to make sure that aid works. Dropping the ball on new methods to evaluate impact suggests that this responsibility is not taken seriously enough. This should be a strong incentive for politicians and high level administrators with a stake in keeping Sweden's good reputation alive.

Finally, but not least importantly, impact evaluation can also promote some discipline in the whole process of planning and defining the objectives of aid activities. As mentioned above, RCTs typically require that the evaluation team is part of the process from scratch, and hoped for outcomes must be well-defined, observable and quantifiable. This can help solve some common complaints when aid

[28] Ironically, talking to academic economists not directly involved in this kind of work you often get the opposite reaction; RCTs sound useful and make sense, but why do you call it research? Where is the theory?

[29] Banerjee and Duflo (2012) ranks on spot 4 on the Amazon top-seller list in the category of International Economics, suggesting that it is possible to create interest among a broader public for a book on development largely based on findings using RCTs.

financed projects are evaluated; unclear objectives, lacking baselines and hard to measure intended outcomes (e.g. Statskontoret 2012). For instance, irrespectively whether it is an impact or effectiveness evaluation, a baseline is typically necessary to be able to judge the situation prior to the intervention, and a control group is essential to get at least an approximate answer to the counterfactual question what the outcome would have been in the absence of the intervention. An RCT design makes sure that these pieces are ticked off. It can also help shift part of the focus from output to impact, thereby remedying the current imbalance in the evaluation portfolio.

Towards a more evidence based aid policy

In this subsection I will take some tentative steps towards suggestions for how to incorporate RCTs and more evidence based thinking into Swedish aid operations. What I have in mind is not to just squeeze in an RCT here and there, but a more fundamental change in which understanding impact should be an ambition for all activities where impact is part of the objective, and where the best available methods should be used to the extent possible. These suggestions should be thought of as the beginning of a discussion of how to use better methods to improve aid effectiveness in Sweden, from the planning to the evaluation stage. How to more concretely adapt this into the specific institutional and organizational setting of Sida and Swedish aid more generally would be the next step (if it is deemed desirable), but that goes beyond the scope of this report.

First there must be recognition from the top that when it comes to evaluation, using credible methods to measure impact is essential,

possible and expected. It is not realistic to believe that desk officers and their counterparts will organically start designing projects and programs with RCTs in mind, without being given the incentives and directives from above. Leadership (within both government and Sida) must also make sure that resources are available to finance RCTs and, equally important, train staff in how to use RCTs in project design and priorities. This includes being able to read, understand and critically assess existing studies, understand how to design and implement a project portfolio to make robust impact evaluation possible, and the knowledge to commission new RCTs when judged appropriate.[30]

To be more specific, the idea is of course not to have an RCT conducted on every project. There are many different ways to take advantage of available knowledge and resources to foster a more evidence based approach to aid policy. The public good character of RCTs means that there is a knowledge bank out there with evidence of what works and what doesn't. The limitations of the methodology, in particular external validity, must of course be kept in mind when making inference from other studies, but the increased availability of replication studies reduces the concern with case specific results. For the purpose of planning activities, and finding priorities, *creating a knowledge bank of existing evaluations and relevant research, including, but not exclusively, RCTs and other impact evaluations, should thus be a first priority*.[31] A very good recent initiative in this direction is Sida's

[30] I am not saying all staff must have this competence, but there needs to be a resource available that can get involved when needed.
[31] Note that this is something completely different than the "Öppna biståndet" initiative (at openaid.se). The transparency offered is praiseworthy, but the focus on tracking money and

support to, and collaboration with, ReCom (Research and Communication on International Aid), a research program under UNU-WIDER. The purpose of ReCom is to gather research and knowledge of relevance for the priorities of Swedish and Danish aid policy into a knowledge bank, along the lines suggested above. Findings are then disseminated through meetings in Stockholm and Copenhagen, and through the programs webpage. This is a promising initiative, and can be very helpful to make decisions on aid more evidence based. In the end, though, whether this potential is realized or not will depend on the extent to which decision makers also use this resource as an input into their decisions.

For this knowledge bank to be useful, officers in charge of course need the skills to read, understand and critically asses the relevance, generalizability, internal validity and comparative strengths and weaknesses of evaluations using different methods. A second priority should thus be to *make sure that relevant staff has the capacity to read and apply the lessons from impact evaluations, and that training is offered if deemed necessary*. The UK Department for International Development (DFID) has made a concerted effort to make their work more evidence based in many different ways.[32] For that purpose the agency has gotten more resources and an increase in hiring of staff with higher degrees. There is little that suggests that Sida will be granted the same opportunities in the near future. However, also

output is also a reflection of how the thinking at times is a bit off what may be the most important matters.

[32] They do for instance have a Research4development on-line portal that offers information about all research being funded by DFID. This can be a model to the knowledge bank I discuss above, but there is no reason to include only research financed by Sida in Sweden's case (much of which I assume is already available on the web).

within a given budget frame there are lessons to learn from DFID. For instance, an interesting (and not expensive to adopt to a Swedish setting) resource is the "Assessing the Strength of Evidence" How to Note, launched in February 2013. The note builds on DFID's ambition to base spending decisions on the best available evidence throughout the organization, and offers a short (21 pages) guide for staff on how to evaluate existing, sometimes conflicting, evidence.[33] The guide brings up both qualitative and quantitative methods, explains how they work and discuss what can, and cannot, be learned from them. It is also made explicit that the note *should* be applied to "Evidence Papers", provided internally in DFID, and that it should be used at least as a guide in all other evidence products.

There will also be situations when Sida needs to commission their own RCTs, though. This can either be when planning a new type of activity and existing information on the relative effectiveness of different approaches is not available, i.e. the long term objective of an impact evaluation discussed above. In this case a pilot study of an initial, maybe small scale, intervention is warranted. Another case is when Sida, or the government, is interested not in the effectiveness of a general type of intervention for planning ahead, but a specific intervention it is already financing. As discussed above, a difference between a researcher and an aid agency or donor country government is that the latter two could be interested in the impact of the specific intervention they have financed simply because it is their project, and

[33] DFID also, among many other things, provide an online guide to research designs and methods, a handbook on research and evaluation methods, and an introduction on how to use statistics (see https://www.gov.uk/government/organisations/department-for-international-development/about/research#research-database).

it is part of their mandate to evaluate whether the specific projects they finance are effective or not (what I have referred to as the short term objective of an impact evaluation). A third priority is thus to *develop in-house competence to commission RCTs*. When should they be commissioned? Who has the competence to perform the RCT? What is a reasonable time frame and budget? How is a terms of reference written, and how are tenders evaluated with regards to quality and feasibility of approach?

A key challenge to making aid policy more evidence based and build in RCT's into the process is the organization of the "value chain" of aid. This has partly to do with the organizational structure within Sida, but it also reaches beyond that and the role of implementing units (NGO's and consultancy firms) and partner countries. As discussed above, an RCT requires randomized assignment of the intervention and that randomization should preferably take place before the intervention is started (the exception is encouragement designs). This means that evaluation needs to be planned for already from the very first stage, and it entails some restrictions on the discretion on how to implement and target the intervention. From an organizational perspective at Sida, this means that the desk officer in charge of a project/program needs to make sure that credible preconditions for impact evaluation are there in the proposed design of the project, and that Sida has the capacity to monitor that the plans for rollout also are followed. There is thus an organizational task in making sure that the capacity is present to judge whether the proposed project satisfies the requirements or not (or how it can be accommodated to do so) in a robust and credible way.

This suggests that a forth priority is that *Sida's unit for monitoring and evaluation (UTV) should be engaged already at the assessment and approval stage of new projects/programs where impact evaluation is deemed feasible and relevant*. Another reorganization may be the last thing Sida needs at this point, but there needs to be a routine in place to make sure that this works effectively.

Beyond the internal organization within Sida, incorporating RCT's into project design requires collaboration from partner country counterparts and agents responsible for implementation. The Paris Declaration and subsequent amendments emphasize results, but also the importance of ownership and alignment with country systems. Donors such as Sida are of course still able to require that projects and programs suggested by partner countries are designed such that proper evaluation of outcomes and impact can be done. However, it does imply that a reasonable fifth priority is that Sida *spend more effort and resources on communication and training of counterparts in what designing projects and programs for impact evaluation requires, and why it is important*.[34]

This is not a trivial task, though. RCTs have become more and more common, and most aid practitioners and key partner country counterparts probably know, at least roughly, what it entails. However, as discussed for instance in Hayman and Bartlett (2013), many practitioners are still struggling with tighter requirements on robust evidence. Challenges include poor access to relevant

[34] This is not to say that Sida is not already engaged in this. Sida is part of financing several initiatives in this area, such as 3ie, CLEAR, and the Big Push Forward network. However, even more may be needed.

information and training, capacity and resource constraints, and a lack of internal understanding of why this is important.[35] RCTs typically require an approach different from how interventions typically have been done, add new requirements on team leaders and task managers, and mean less control and discretion on behalf of the implementing agency. There is a fair degree of skepticism within the community of practitioners, as expressed for instance by the "Big Push Forward" network (http://bigpushforward.net), towards the general "results and evidence agenda".[36] There is a struggle with mapping the evidence to action, the relevance of studies in different environments is questioned (serious doubts about external validity), the evidence agenda is sometimes seen as political and hand-in-hand with a marketization of aid, and there are ethical concerns with the whole approach within NGOs struggling to explain to control group members why they have been excluded from the intervention.

Some of this skepticism and critique is well taken and part of the back and forth towards finding the right balance between control and discretion, and identifying ways of learning from different methods and experience. Nevertheless, most practitioners can also see benefits of the approach when applied in the right context. It is also important to note that incentives of NGOs are not always aligned with those of policy makers and tax payers in donor countries. As in most

[35] WHO and DFID in particular have been quite active in capacity building initiatives and communication strategies to promote the use of research and evidence in the field. This has primarily targeted policy makers and government officials, though, while NGOs have received much less attention (Hayman and Bartlett, 2013).

[36] This "agenda" stretches far beyond just RCTs and impact evaluation. It is defined as follows in The Big Push Forward (2013, p. 1): "The pursuit of information on (intended) results and evidence of results to justify aid, improve aid and manage aid agencies through protocols, procedures and mechanisms for reporting, tracking, disbursement mechanisms, appraising, and evaluating effectiveness and impact."

organizations there is often a resistance to change motivated as much by habit formation as anything else, and it is natural to prefer to keep discretion in terms of whom to target and how to operate. Having operations evaluated does of course also carry the risk that results come up negative. For NGOs who have succeeded in securing steady streams of finance, the potential upside of an impact evaluation is not that big (donors already seem to believe, correctly or not, that the NGO makes a difference), whereas the potential downside could be huge if results turn up negative. Saying that everything is win-win would be to trivialize the challenges involved. Both carrots and sticks may be necessary to get all parties onboard, and, most importantly, donor agencies such as Sida need to be ready to invest resources into communication, training and persuasion.[37]

An interesting initiative, that Sida supports, is CLEAR (Regional Centers for Learning and Results). This is a collaborative effort between donors and partner countries that aim to build local competence in monitoring, evaluation and performance evaluation. Regional academic centers, chosen competitively, serve as hubs of experience exchange and knowledge, and they organize workshops and training. Initiatives such as these serve at least two purposes. A direct development assistance effect by contributing to human capital development and the institutional capacity in partner countries. But, also an indirect effect of creating a potential pool of partner country competence that can be used to conduct robust impact evaluation, if properly trained. In addition to the "aid externality", tapping into this

[37] That the relationship between Sida and contracting NGOs are not always a win-win situation is also illustrated by the effect of formal monitoring as illustrated in the aforementioned paper by Bengtsson and Engström (2013).

resource also has the advantages that those responsible should have the necessary knowledge to understand and be operational in the local context, and it should be relatively cheap. Supporting initiatives such as this seems to be an opportunity to generate many positive effects at once, though quality control of course will be essential.

Finally, I also want to bring up that there are indirect ways through which Sida and Sweden can contribute to the public good of knowledge about what works and not in development, beyond commissioning evaluations of their own. There are research networks such as J-PAL at MIT and Innovations for Poverty Action (IPA) at Yale University that support randomized evaluations primarily by researchers, but that also have as explicit objectives that findings should lead to policy action. Another initiative beyond the strict scholarly community is 3ie (International Initiative for Impact Evaluation).[38] Set up as an international collaborative effort to promote the use of impact evaluation for policy purposes, 3ie offers resources to undertake impact evaluation and to disseminate the findings to policy makers. The emphasis of policy impact means among other things that the organization explicitly strive to "…contribute evidence that leverages information from other sources and studies" (3ie 2008, page 3). Thus, contrary to the academic community, that puts a very high value on novelty and originality, and thereby somewhat shuns replication studies that may have high policy relevance, 3ie explicitly encourages this. An indirect way for Sida to support the creation of evidence and use of impact evaluation is thus

[38] 3ie does not explicitly require RCTs, but it defines rigorous impact evaluation as using the best methodology available, which in practice typically means an RCT.

to contribute to organizations such as these. This is also already done to some extent, but it is an open question if it is motivated to increase such support.

Conclusions

The purpose of this report was twofold. First to bring up how Randomized Controlled Trials (RCTs) have been used in research and practice to promote evidence based development policy. Second to discuss to what extent, up until now, this tool has been used by Swedish aid practitioners, if there are reasons to expand that use, and if so, offer some tentative suggestion for how that can be achieved.

The main arguments of the report are that:

- RCTs is a very powerful and useful method to evaluate impact under the correct circumstances, and can help decision makers better allocate resources towards interventions that make a real difference in the life of aid recipients. It should therefore be part of the toolbox of aid agencies that have an obligation to make sure that development finance is allocated towards interventions that work.
- RCTs also have significant limitations, though, and can neither generate answers to all policy relevant questions, nor be applied to all types of projects and programs. It can thus be no more than one of many tools for monitoring and evaluation used by aid agencies. Using mixed methods, a combination of RCTs and other quantitative or qualitative methods, will often be the ideal approach to learn more broadly about the effectiveness of different dimensions of aid financed interventions.
- The use of RCTs has moved beyond the academic community and has become fairly common among the more progressive

multilateral and bilateral donors. Very little of this has been seen in Sweden. It is somewhat surprising that a country with such a high profile in this area of foreign policy, a generally solid reputation within the aid community, and a publicly stated emphasis on results, has shown so little enthusiasm to use this tool to evaluate the projects and programs it finances.

- The advantages of including RCTs into Swedish aid practices also go beyond just getting a better understanding of impact. First, it would lend more credibility to the ambition to be a serious and unbiased partner in the international aid community. Second, RCTs require a firm understanding of what exactly the objectives of the intervention are, and how their fulfilment can be measured, already from the start. This helps avoid common pitfalls with aid financed interventions; unclear objectives, unobservable or unmeasurable intended outcomes, and the inability to even quantify changes in outcomes in the targeted group due to missing baselines.

- A step towards more evidence based planning of projects and programs is to start using the existing bank of knowledge that exists in the form of already done impact evaluations. This requires that staff have access to existing material, know how to read and evaluate evidence derived using different methods, and know how to address limitations with regards to for instance external validity. This may require training and access to advice and help from staff specialized in monitoring and evaluation. The ReCom initiative is a very good step in that direction.

- Another key step is to acquire the in-house competence to commission RCTs of projects and programs financed by Sida or collaborators. This evaluation competence should then ideally be involved already at the initiation stage of a project (not at an intermediate stage, or the final stage, as is typically done now), as RCTs typically impose requirements on how the intervention is conducted from the very start. This also secures the benefits alluded to above, in terms of well-defined objectives and measurable anticipated outcomes.
- Finally, conducting RCTs require the collaboration of partner countries and implementing units (NGOs or consultancy firms). This may require collaborative training efforts, and at times some convincing. A possibly fruitful approach to combine partner country human capital development with the creation of a resource for conducting RCTs is to support development of impact evaluation skills at partner country universities and research centres.

I can only speculate about why the Swedish aid community has been relatively slow on picking up this methodology. As often, there is probably a combination of factors that matter, and the last few years turmoil with budget conflicts, management changes and reorganizations has probably focused attention towards more short term objectives. Nevertheless, as has been pointed out before (e.g. Sida 2008), there has never really been much attention paid to any type of quantitative methods of evaluation at Sida. It is easy to get the impression that the very reasonable theoretical objection that

quantitative methods and RCTs cannot be used everywhere, or answer all questions, in practice has led to a policy where these methods are used nowhere, and to answer no questions. When it comes to impact evaluation, I think this is partly a reflection of misconceptions about what RCTs do and don't. The first concern to address is the perception that RCTs are research, and therefore irrelevant for policy purposes. Clearly what should matter for policy makers and their decisions is information about the impact of aid financed projects on the intended recipients. In the right circumstances, an RCT is generally regarded as the best method for that purpose. It is not a highly theoretical or sophisticated concept that relies on some obscure set of underlying assumptions. Rather the opposite, the ambition is generally to make it as independent as possible from behavioural assumptions. And, even though the methodology cannot answer all questions and be applied to all interventions, it is still quite versatile and is continuously being developed to tackle some of its weaknesses, such as external validity.

A second concern relates to the resource requirements that come with a longer planning horizon and increased financial costs. There is definitively some truth to this, but it is also partly a fallacy of thinking because with the right planning the initial phase (the baseline survey and the randomization) should rather be thought of as part of preparation when a new project is implemented. Conducting and analysing the end-line survey should not be much more time

consuming or expensive than alternative methods of evaluation.[39] The binding constraint is thus not so much time or money but rather the planning horizon and the ability to coordinate evaluation and implementation already from the get-go. As pointed out above, this could also facilitate with some of the other limitations often faced by evaluation teams, such as unclear objectives, non-measurable target outcomes, and lack of documentation of key outcome variables. The methodology thus imposes a certain degree of discipline into the process of initiating new projects and programs. This may not be a bad idea in an activity sometimes focusing too much on getting money and projects through the door.

Looking forward, this report can hopefully serve as an input to the discussion on Swedish aid policy, but to move from discussion to action, more is needed. As suggested here, taking impact evaluation more seriously has some implications for how to organize work internally as well as externally towards partner countries and implementing NGOs and consultancy firms. Some of this can be learned from the experience of other aid agencies that have made more progress towards an evidence based approach to planning and implementation, but there must also be an analysis of how to fit this into the reality of Swedish aid specifically. An in-depth organizational study of how to practically organize the processes through which projects and programs are selected, designed and implemented, with the possibility of rigorous impact evaluation in mind, would thus be an

[39] The costs of more rigorous evaluation methods are also likely to be smaller than the costs of repeatedly implementing the wrong interventions due to lack of knowledge of what works and what doesn't.

important concrete step forward towards a more evidence based approach to Swedish development co-operation.

References

3ie (2008), “Founding Document for Establishing the International Initiative for Impact Evaluation”, available at http://www.3ieimpact.org/media/filer/2012/05/17/3iefoundingdocument30june2008.pdf.

Acemoglu, D. (2010), "Theory, General Equilibrium and Political Economy in Development Economics", Journal of Economic Perspectives 24, 17-32.

Angelucci, M., D. Karlan, and J. Zinman (2012), “Win some lose some? Evidence from a randomized microcredit program placement experiment by Compartamos Banco”, J-PAL working paper.

Attanasio, O., B. Augsburg, R. De Haas, E. Fitzsimons, and H. Harmgart (2011), “Group lending or individual lending? Evidence from a randomised field experiment in Mongolia”, MPRA Paper No. 35439.

Augsburg, B., R. D. Haas, H. Harmgart, and C. Meghir (2012), “Microfinance, poverty and education”, IFS working paper.

Baird, S., C. McIntosh and B. Özler (2011), “Cash or Condition? Evidence from a Randomized Cash Transfer Program”, Quarterly Journal of Economics 126(4): 1709-1753.

Bamberger, M., V. Rao, and M. Woolcock (2010), “Using Mixed Methods in Monitoring and Evaluation”, Policy Research Working Paper 5245, World Bank, Washington DC.

Banerjee, A. (2007), "Making aid work", The MIT Press, Cambridge, Massachusetts.

Banerjee, A and E. Duflo (2009), "The Experimental Approach to Development Economics", Annual Review of Economics 1, 151.78.

Banerjee, A., S. Kumar, R. Pande, and F. Su (2011), "Do Informed Voters make better choices? Experimental Evidence from Rural India", mimeo, MIT.

Banerjee, A and E. Duflo (2012), "Poor Economics: A Radical Rethinking of the Way to Fight Poverty", MIT Press.

Banerjee, A., E. Duflo, R. Glennerster and C. Kinnan (2013), "The Miracle of Microfinance? Evidence from a Randomized Evaluation", Working Paper, MIT.

Bengtsson, N. and P. Engström (2013), "Replacing Trust with Control: A Field Test of Motivation Crowd out Theory", The Economic Journal , forthcoming.

Benhassine, N., F. Devoto, E. Duflo, P. Dupas and V. Pouliquen (2010), "Turning a Shove Into a Nudge? A "Labeled Cash Transfer" for Education", Working Paper, Stanford University.

Björkman-Nyqvist, M., and J. Svensson (2009), "Power to the People: Evidence from a Randomized Experiment on Community-Based Monitoring in Uganda", Quarterly Journal of Economics, 124:2: 735–769.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a and J. Sandefur (2013), "Scaling up What Works: Experimental Evidence on External

Validity in Kenyan Education", mimeo, IIES, Stockholm University.

Brigham, M., M. Findley, W. Matthias, C. Petrey, and D. Nelson (2013), "Aversion to Learning in Development? A Global Field Experiment on Microfinance Institutions", Technical Report, Brigham Young University.

Bulte, E., L. Pan, J. Hella, G. Beekman and S. di Falco (2012). Pseudo-Placebo Effects in Randomized Controlled Trials for Development: Evidence from a Double-Blind Field Experiment in Tanzania. Working Paper.

Crépon, B., E. Duflo, F. Devoto, and W. Pariente (2011), "Impact of microcredit in rural areas of Morocco: Evidence from a randomized evaluation", J-PAL working paper.

Deaton, A. (2009), "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive keys to Economic Development", NBER Working Paper 14690.

Desai, R. and S. Joshi (2014), "Collective Action and Community Development: Evidence from Self-Help Groups in Rural India", The World Bank Economic Review.

Desai, R., S. Joshi and A. Olofsgard (2014), "Can the Poor Be Organized? Behavioral Evidence from Rural India", mimeo.

DFID (2012), "Broadening the Range of Designs and Methods for Impact Evaluation", DFID Working Paper 38.

DFID (2013), "Assessing the Strength of Evidence", DFID How to Note, UK.

Dickson, R., S. Awasthi, P. Williamson, C. Demellweek, and P. Garner (2000), "Effect of treatment for intestinal helminthes infection on growth and cognitive performance in children: systematic review of randomized trials", British Medical Journal, 320 (June 24), 1697-1701.

Duflo, E. and R. Chattopadhyay (2004), "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India", Econometrica 72, 1409-1443.

Duflo, E., P. Dupas and M. Kremer (2008), "Peer Effects, Pupil Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya," Working Paper, MIT.

Hayman, R. and J. Bartlett (2013), "Getting to Grips with Evidence; How NGOs can Tackle changing needs in the use of Evidence and Research", Praxis Paper 28, INTRAC.

Heckman, J. and J. Smith (1995), "Assessing the Case for Social Experiments", Journal of Economic Perspectives, 9, 85-110.

Karlan, D. and J. Zinman (2011), "Microcredit in theory and practice: Using randomized credit scoring for impact evaluation", Science 332 (6035), 1278–1284.

Martens, B. (2002), "The Role of Evaluation in Foreign Aid Programmes", in Martens et al. The Institutional Economics of Foreign Aid, Cambridge University Press.

Miguel, E. and M. Kremer (2004), "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," Econometrica, Volume72 (1), pp. 159-217.

Natsios, A. (2010), "The Clash of the Counter-bureaucracy and Development", CGD Essay, July 2010.

OECD/DAC (2002), "Development Assistance Manual", OECD, Paris.

Olken, B. (2007), "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," Journal of Political Economy, Volume 115 (2), pp. 200-249.

Olofsgard, A. (2012), "The Politics of Aid Effectiveness: Why Better Tools can make for Worse Outcomes", SITE WP No. 16.

Pritchett, L. and J. Sandefur (2014), "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix", Journal of Globalization and Development 4, Issue 2, 161–197

Ravallion, M. (2009), "Should the Randomistas Rule?", Economists Voice 6.

Rodrik, D. (2009), "The New Development Economics: We Shall Experiment, but How Shall We Learn?". In What Works in Development: Thinking Big and Thinking Small, edited by Jessica Cohen and William Easterly, 24.47. Washington, D.C.: Brookings Institution Press.

Schultz, P. T. (2004), "School subsidies for the poor: evaluating the Mexican Progresa poverty program," Journal of Development Economics, 74(1), 199-250.

Sida (2007), "Looking Back, Moving Forward: Sida Evaluation Manual", Sida, Stockholm.

Sida (2008), "Are Sida Evaluations Good Enough? An Assessment of 34 Evaluation Reports", Sida Studies in Evaluation 2008:1, Stockholm.

Statskontoret (2012), "Utvärdering av Svenskt Bistånd: En Översyn av Utvärderingsverksamheten", Diarienr 2011/250-5.

The Big Push Forward (2013), "The Politics of Evidence Conference Report", Institute of Development Studies, UK.

Transtec (2014), "Mid Term Review of the Framework Agreement for Sida Reviews, Evaluations and Advisory Services on Results Frameworks", report to Sida, Stockholm.