Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity

Jörg Peters
Jörg Langbein
Gareth Roberts
Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/86788835
ISSN 1864-4872 (online)
ISBN 978-3-86788-835-6
Abstract

When properly implemented, Randomized Controlled Trials (RCT) achieve a high degree of internal validity. Yet, if an RCT is to inform policy, it is critical to establish external validity. This paper systematically reviews all RCTs conducted in developing countries and published in leading economic journals between 2009 and 2014 with respect to how they deal with external validity. Following Duflo, Glennerster, and Kremer (2008), the hazards to external validity we scrutinize are Hawthorne effects, general equilibrium effects, specific sample problems, and special care in treatment provision. Based on a set of objective indicators, we find that the majority of published RCTs does not discuss these hazards and many do not provide the necessary information to assess potential problems. The paper calls for including external validity dimensions in a more systematic reporting on the results of RCTs. This may create incentives to avoid overgeneralizing findings and help policy makers to interpret results appropriately.

JEL Classification: C83, C93

Keywords: Policy evaluation; systematic review; internal validity; external validity; randomized controlled trials

October 2017
1. Introduction

In recent years, an intense debate has taken place about the value of Randomized Controlled Trials (RCTs). Most notably in development economics, RCTs have assumed a dominating role. The striking advantage of RCTs is that they overcome self-selection into treatment and thus their internal validity is indisputably high. This merit is sometimes contrasted with shortcomings in external validity (Basu 2014; Deaton and Cartwright 2016). Critics state that establishing external validity is more difficult for RCTs than for studies based on observational data (Dehejia 2015, Muller 2015, Moffit 2004, Pritchett and Sandefur 2015, Roe and Just 2009, and Temple 2010). This is particularly true for RCTs in the development context that tend to be implemented at smaller scale and in a specific locality. Scaling an intervention is likely to change the treatment effects, because the scaled program is typically implemented by resource-constrained governments, while the original RCT is often implemented by effective NGOs or the researchers themselves (Banerjee et al. 2016, Bold et al. 2013, Deaton and Cartwright 2016, and Ravallion 2012).

This does not question the enormous contribution that RCTs have made to the knowledge about the effectiveness of policy interventions. It rather underscores that “research designs in economics offer no free lunches – no single approach universally solves problems of general validity without imposing other limitations” (Roe and Just 2009). Indeed, Rodrik (2009) argues that RCTs require “credibility-enhancing arguments” to support their external validity – just as observational studies have to make a stronger case for internal validity. Against this background, the present paper examines how the results published from RCT based evaluations are reported, whether external validity relevant design features are made transparent, and whether potential limitations to transferability are discussed.

To this end, we conduct a systematic review of policy evaluations based on RCTs published in top economic journals. We include all RCTs published between 2009 and 2014 in the American Economic Review, the Quarterly Journal of Economics, Econometrica, the Economic Journal, the Review of Economic Studies, the Review of Economics and
Statistics, the *Journal of Political Economy* and the *American Economic Journal: Applied Economics*. In total, we identified 54 RCT-based papers in these journals.

Since there is no uniform definition of external validity and its hazards in the literature, in a first step we establish a theoretical framework deducing the assumptions required to transfer findings from an RCT to another policy population. We do this based on a model from the philosophical literature on the probabilistic theory of causality provided by CARTWRIGHT (2010) and based on a seminal contribution to the economics literature, the toolkit for the implementation of RCTs by DUFLO, GLENNERSTER, AND KREMER (2008). We identify four hazards to external validity: i) Hawthorne and John Henry Effects, ii) general equilibrium effects, iii) specific sample problems, and iv) problems that occur when the treatment in the RCT is provided with special care compared to how it would be implemented under real-world conditions.

As a second step, we scrutinize the reviewed papers with regard to how they deal with the four external validity dimensions and whether required assumptions are discussed. Along the lines of these hazards we formulated seven questions, then read all 54 papers carefully and asked each of them these seven questions. All questions can be objectively answered by ‘yes’ or ‘no’; no subjective rating is involved.

External validity is not necessary in some cases. For example, when RCTs are used for accountability reasons by a donor or a government, the results are only interpreted within the evaluated population. Yet, as soon as these findings are used to inform policy elsewhere or at larger scale, external validity becomes a pivotal element. Moreover, test-of-a-theory or proof of concept RCTs that set out to disprove a general theoretical proposition speak for themselves and do not need to establish external validity (DEATON AND CARTWRIGHT 2016). However, in academic research most RCTs presumably intend to inform policy and, as we will also confirm in the review, the vast majority of included papers appear to generalize findings from the study population to a different policy population.²

Indeed, RCT proponents in the development community advocate in favor of RCTs in order to create “global public goods” that “can offer reliable guidance to international
organizations, governments, donors, and NGOs beyond national borders” (DUFLO, GLENNERSTER AND KREMER 2008, p. 93). Already in 2005, during a symposium on “New directions in development economics: Theory or empirics?” Abhijit Banerjee acknowledged the requirement to establish external validity for RCTs and, like Rodrik, called for arguments that establish the external validity of RCTs (BANERJEE 2005). To conclude, Banerjee and Rodrik seem to agree that external validity is never a self-evident fact in empirical research and that RCTs in particular have to discuss in how far results are generalizable.

In the remainder of the paper we first present the theoretical framework and establish the four hazards to external validity (Section 1), before the methodological approach and the seven questions are discussed in Section 2. The results are presented in Section 3 and discussed in Section 4. Section 5 provides an overview on existing remedies for external validity problems and ways to deal with them in practice. Section 6 concludes.

2. Theoretical Background and Definition of External Validity

2.1 Theoretical Framework

The understanding of what external validity exactly is and how it might be threatened is not clearly defined in the literature. What we are interested in here is the degree to which an internally valid finding obtained in an RCT is relevant for policy makers who want to implement the same intervention in a different policy population. CARTWRIGHT (2010, p. 60) defines external validity in a way that is similar to the understanding conveyed in DUFLO, GLENNERSTER AND KREMER (2008, DGK in the following): “External validity has to do with whether the result that is established in the study will be true elsewhere.” Cartwright provides a model based on the probabilistic theory of causality. Using this model we identify the assumptions that have to be made when transferring the results from an RCT to what a policy maker can expect if she scales the intervention under real-world conditions.
Suppose we are interested in whether a policy intervention $C$ affects a certain outcome $E$, we can state that $C$ causes $E$ if

$$P(E|C & K_i) > P(E|\bar{C} & K_i)$$

where $K_i$ describes the environment and intervention particularities under which the observation is made. Assume this causal relationship was observed in population $A$ and we want to transfer it to a situation in which $C$ is introduced to another population $A'$. In this case, Cartwright points out that those observations $K_i$ have to be identical in both populations $A$ and $A'$ as soon as they interfere with the treatment effect. More specifically, Cartwright formulates the following assumptions that are required:

(i) $A$ needs to be a representative sample of $A'$

(ii) $C$ is introduced in $A'$ as it was in the experiment in $A$

(iii) The introduction leaves the causal structure in $A'$ unchanged.

In the following, we use the language that is widely used in the economics literature and refer to the toolkit for the implementation of RCTs by DGK (2008). Similar to the Cartwright framework, DGK introduce external validity as the question “[…] whether the impact we measure would carry over to other samples or populations. In other words, whether the results are generalizable and replicable” (p. 3950). The four hazards to external validity that are identified by DGK are Hawthorne and John Henry Effects, general equilibrium effects, the specific sample problem and the special care problem. The following section presents these hazards to external validity in more detail. Under the assumption that observational studies mostly evaluate policy interventions that would have been implemented anyhow, Hawthorne/John Henry Effects and the special care problem are much more likely in RCTs, while general equilibrium effects and the specific sample problem equally occur in RCTs and observational studies.
2.2 Potential Hazards to External Validity

In order to guide the introduction to the different hazards of external validity we use a stylized intervention of a cash transfer given to young adults in an African village. Suppose the transfer is randomly assigned among young male adults in the village. The evaluation examines the consumption patterns of the recipients. We observe that the transfer receivers use the money to buy some food for their families, football shirts, and air time for their mobile phones. In comparison, those villagers, who did not receive the transfer, will not change their consumption patterns. What would this observation tell us about giving a cash transfer to people in different set-ups? The answer to this question depends on the assumptions identified in DGKs’ nomenclature.

Hawthorne and John Henry effects might occur if the participants in an RCT know or notice that they are part of an experiment and under observation.\(^3\) It is obvious that this could lead to an altered behavior in the treatment group (Hawthorne effect) and/or the control group (John Henry effect).\(^4\) In the stylized cash transfer example the recipient of the transfer can be expected to spend the money for other purposes in case he knows that his behavior is under observation. It is also obvious that such behavioral responses clearly differ between different experimental set-ups. If the experiment is embedded into a business-as-usual set up, distortions of participants’ behavior are less likely. In contrast, if the randomized intervention interferes noticeably with the participants’ daily life (for example, an NGO appearing in an African village to randomize a certain training measure among the villagers), participants will probably behave differently than they would under non-experimental conditions.\(^5\)

The special care problem refers to the fact that in RCTs the treatment is provided differently from what would be done in a non-controlled program. In the stylized cash transfer example, a lump sum payment that is scaled up would perhaps be provided by a larger implementing agency with less personal contact. Bold et al. (2013) provide compelling evidence for the special care-effect in an RCT that was scaled up based on positive effects observed in a smaller RCT conducted by Duflo et al. (2011b).
major difference is that the program examined in Bold et al. was implemented by the national government instead of an NGO in the Duflo et al. study. The positive results observed in Duflo et al. could not be replicated in Bold et al. According to the authors:

“Our results suggest that scaling-up an intervention (typically defined at the school, clinic, or village level) found to work in a randomized trial run by a specific organization (often an NGO chosen for its organizational efficiency) requires an understanding of the whole delivery chain. If this delivery chain involves a government Ministry with limited implementation capacity or which is subject to considerable political pressures, agents may respond differently than they would to an NGO-led experiment” (p. 29f.).

VIVALT (2016) confirms the higher effectiveness of RCTs implemented by NGOs or the researchers themselves as compared to RCTs implemented by governments in a meta-analysis of published RCTs. Further evidence on the special care problem is provided by ALLCOTT (2015). He shows that electricity providers that implemented RCTs in cooperation with a large research program to evaluate household energy conservation instruments are systematically different from those electricity providers that do not participate in this program. This hints at what Allcott refer to as “site selection bias”: Organizations that agree to cooperate with researchers on an RCT can be expected to be different compared to those that do not, for example because their staff are more motivated. This difference could translate into a higher general effectiveness. Therefore, the effectiveness observed in RCTs is probably higher than it will be when the evaluated program is scaled to those organizations that did not cooperate with researchers at first.

The third identified hazard arises from potential **general equilibrium effects (GEE)**. Typically, such GEE only become noticeable if the program is scaled to a broader population or extended to a longer term. In the stylized cash transfer example provided above, GEE occur if not only a small number of people but many villagers receive the transfer payment. In this scaled version of the intervention, some of the products that young male villagers want to buy become scarcer and, thus, more expensive. This also illustrates that GEE can affect non-treated villagers, as prices increase for them as well. Moreover, in the longer term if the cash transfer program is
implemented permanently, certain norms and attitudes towards labor supply or educational investment might change.\textsuperscript{7}

This example indicates that GEE in their entirety are difficult to capture. The severity of GEE, though, depends on some parameters like the regional coverage of the RCT, the time horizon of the measurements, and the impact indicators the study looks at. Very small-scale RCTs or those that measure outcomes after a few months only are unlikely to portray the change in norms and beliefs that the intervention might entail. Furthermore, especially market based outcomes like wages or employment status will certainly be affected by adjustments in the general equilibrium if an intervention is scaled and implemented over many years. As a matter of course, it is beyond the scope of most studies to comprehensively account for such GEE and RCTs that cleanly identify partial equilibrium effects can still be informative for policy. A profound discussion of GEE relevant features is nonetheless necessary to avoid the ill-advised interpretation of results. Note that GEE are not particular to RCTs and, all else equal, the generalizability of the results from observational studies is also exposed by potential GEE. Many RCTs, in particular in developing country contexts, are however limited to a specific region, a relatively small sample size, and short monitoring horizon and are thus more prone to GEE than country-wide representative panel-data based observational studies.

In a similar vein, the fourth hazard to external validity, the \textbf{specific sample problem}, is not particular to RCTs but might be more pronounced in this setting. The problem occurs if the study population is different from the policy population in which the intervention will be brought to scale. Taking the cash transfer example, the treatment effect for young male adults can be expected to be different if the cash transfer is given to young female adults in the same village or to young male adults in a different part of the country.

In Section 2.2, these hazards to external validity are translated into objective questions to be asked during the systematic review of published RCTs.
3. Methods and Data

3.1 Review approach

We reviewed all RCTs conducted in developing countries and published between 2009 and 2014 in the leading journals in economics. We included the five most important economics journals, namely the *American Economic Review*, *Econometrica*, the *Quarterly Journal of Economics*, the *Journal of Political Economy*, and the *Review of Economic Studies*. In addition, we included further leading general interest journals that publish empirical work and RCTs in particular: The *American Economic Journal: Applied Economics*, the *Economic Journal*, and the *Review of Economics and Statistics*.

We scrutinized all issues in the period; all papers that mention either the terms “field experiment”, “randomized controlled trials” or “experimental evidence” in either the title or the abstract or that indicated in the abstract or the title that a policy intervention was randomly introduced were examined further. We excluded those papers that examine interventions in an OECD member country. In total, 73 papers were initially identified. Our focus is on policy evaluation and we therefore excluded mere test-of-a-theory papers. In most cases, the demarcation was very obvious and we subsequently excluded 19 papers. In total, we found 54 papers based on an RCT to evaluate a certain policy intervention in a developing country. The distribution across journals is uneven with the vast majority being published in the *American Economic Journal: Applied Economics*, the *American Economic Review* and the *Quarterly Journal of Economics* (see Figure 1).

Figure 2 depicts the regional coverage of the surveyed RCTs. The number of RCTs implemented in Kenya is due to the strong connection that two of the most prominent organizations that conduct RCTs have to the country (Innovation for Poverty Action, IPA, and the Abdul Latif Jameel Poverty Action Lab J-Pal). Most of their studies were implemented in Kenya’s Western Province by the Dutch NGO International Child Support (ICS), IPA and J-Pal’s cooperation partner in the country.
Figure 1: Published RCTs between 2009 and 2014 (54 studies included in total, frequencies in bold)

Figure 2: Countries of implementation (54 studies included in total, frequencies in bold)
We read all 54 papers carefully (including the online appendix) and each paper was asked seven objective yes/no-questions. An additional filter question addresses whether the paper has the ambition to generalize. This is necessary, because it is sometimes argued that not all RCTs intend to generate generalizable results and are rather designed to test a theoretical concept. In fact, 96 percent of included papers do generalize (see next section for details on the coding of this question). This is no surprise, since on purpose we excluded test-of-a-theory papers and focused on policy evaluations. The remaining seven questions all address the four hazards to external validity outlined in Section 1 and examine whether the ‘credibility-enhancing arguments’ (RODRIK 2009) are provided to underpin the plausibility of external validity. Appendix A in the Annex shows the answers to the seven questions for all surveyed papers individually. In general, we answered the questions conservatively, i.e. when in doubt we answered in favor of the paper. We abstained from applying subjective ratings in order to avoid room for arbitrariness. A simple report on each paper documents the answers to the seven questions and the quote from the paper underlying the respective answer. We sent these reports out to the lead authors of the included papers and asked them to review our answers for their paper(s). For 36 of the 54 papers we received feedback, based on which we changed an answer from ‘no’ to ‘yes’ in 9 cases (out of 378 questions and answers in total). The comments we received from the authors are included in the reports, if necessary followed by a short reply. The revised reports were sent again to the authors for their information and can be found in the online appendix to this paper.

3.2 Seven questions
To elicit the extent the paper accounts for Hawthorne and John Henry effects we first asked the following objective questions:

1. Does the paper explicitly say whether participants are aware (or not) of being part of an experiment or a study?
This question accounts for whether a paper provides the minimum information that is required to assess whether Hawthorne and John-Henry effects might occur. More would be desirable: In order to make a substantiated assessment of Hawthorne-like distortions, information on the implementation of the experiment, the way how participants were contacted, which specific explanations they received, and the extent to which they were aware of an experiment should be presented. We assume (and confirmed in the review) that papers that receive a ‘no’ for Question 1 do not discuss these issues, because a statement on the participants’ awareness of the study is the obvious point of departure for this discussion. It is important to note that unlike laboratory or medical experiments participants in social science RCTs are not always aware of their participation in an experiment.

Only for those papers that receive a ‘yes’ to question 1 we additionally raise the question:

2. If people are aware of being part of an experiment or a study, does the paper (try to) account for Hawthorne or John-Henry effects (1. in the design of the study, 2. in the interpretation of the treatment/mechanisms, 3. in the interpretation of the size of the impact)?

The next set of questions probes into general equilibrium effects. As outlined in Section 1, we define general equilibrium effects as changes due to an intervention that occur in a noticeable way only if the intervention is scaled or after a longer time period. Two questions capture the two transmission channels via which GEE might materialize:

3. Does the paper explicitly discuss what might happen if the program is scaled up?
4. Does the paper explicitly discuss if and how the treatment effect might change in the long run?13

For both questions, we give the answer ‘yes’ as soon as the respective issue is mentioned in the paper, irrespective of whether we consider the discussion to be comprehensive.
The third hazard is what DGK call the *specific sample problem* and is addressed by Question 5:

5. Does the paper explicitly discuss the policy population (to which the findings are generalized) or potential restrictions in generalizing results from the study population?

We applied this question only to those papers that explicitly generalize beyond the study population (see filter question below and in Section 2.1). As soon as a paper discusses the study population vis-à-vis the policy population, we answered the question with ‘yes’, irrespective of our personal judgement on whether we deem the statement to be plausible and the discussion to be comprehensive.

The fourth hazard, special care, is accounted for by two questions:

6. Does the paper discuss particularities of how the randomized treatment was provided in demarcation to a (potential) real-world intervention?

As soon as the paper makes a statement on the design of the treatment compared to the potential real-world treatment, we answered the question with ‘yes’, again irrespective of our personal judgement of whether we deem the statement to be plausible and comprehensive.

In addition, to account for the concern that RCTs implemented by NGOs or researchers themselves might be more effective than scaled programs implemented by, for example, government agencies, we elicit:

7. Who is the implementation partner of the RCT?

The specific wording of the additional filter question is “Does the paper generalize beyond the study population?”. Our coding of this question certainly leaves more room for ambiguity than the coding for the previous objective questions. We therefore
answered this additional question by a ‘yes’ as soon as the paper makes any generalizing statements (most papers do that in the conclusions) that a mere test-of-a-theory would not make. Note that in this question we do not assess the degree to which the paper generalizes (which in fact varies considerably), but only if it generalizes at all.

4. Results

Table 1 shows the results for the seven questions asked to every paper. As noted above, 96 percent of the reviewed papers generalize their results. This underpins the proposal that these studies should provide ‘credibility enhancing arguments’. It is particularly striking that only 35 percent of the published papers mention whether people are aware of being part of an experiment (Question 1). This number also reveals that it is far from common practice in the economics literature to publish the protocol of the experiment or the communication with the participants. Some papers even mention letters that were sent or read to participants but do not include the content in the main text or the appendix.

Only 46 percent of all papers discuss how effects might change in the longer term and whether some sort of adjustments might occur (Question 4). Here, it is important to note that around 65 percent of the reviewed papers examine impacts less than two years after the randomized treatment; on average impacts are evaluated 17 months after the treatment (not shown in the table). While this is in most cases probably inevitable for practical reasons, a discussion of whether treatment effects might change in the long run, for example based on qualitative evidence or theoretical considerations, would be desirable. Note that most of the papers that do discuss long-term effects are those that in fact look at such long-term effects. In other words, only a small minority of papers that only look at very short term effects provides a discussion of potential changes in the long run.

Likewise, potential changes in treatment effects in case the intervention is scaled are hardly discussed (Question 3, 44 percent of papers); 35 percent of the papers do not mention GEE related issues at all, i.e. received a ‘no’ for Question 3 and 4 (not shown
in Table 1). The best score is achieved for the specific sample problem: 77 percent of papers discuss the policy population or potential restrictions to generalizability.

Table 1: Reporting on external validity in published RCTs

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer is yes (in percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hawthorne and John-Henry Effect</strong></td>
<td></td>
</tr>
<tr>
<td>1. Does the paper explicitly say whether participants are aware of being part of an experiment or a study?</td>
<td>35</td>
</tr>
<tr>
<td>2. Does the paper (try) to account for Hawthorne or John Henry effects?*</td>
<td>29</td>
</tr>
<tr>
<td><strong>General Equilibrium Effects</strong></td>
<td></td>
</tr>
<tr>
<td>3. Discusses what happens if program is scaled up?</td>
<td>44</td>
</tr>
<tr>
<td>4. Discusses changes to treatment effects in the long run?</td>
<td>46</td>
</tr>
<tr>
<td><strong>Specific Sample Problems</strong></td>
<td></td>
</tr>
<tr>
<td>5. Discuss the policy population or potential restrictions to generalizability? †</td>
<td>77</td>
</tr>
<tr>
<td><strong>Special Care</strong></td>
<td></td>
</tr>
<tr>
<td>6. Particularities of how the randomized treatment was provided in demarcation to a (potential) real-world intervention discussed?</td>
<td>20</td>
</tr>
</tbody>
</table>

* Question 3 only applies to those 19 papers that explicitly state that participants are aware of being part of an experiment. † Question 5 only applies to those 52 papers that explicitly generalize.

As the results for Question 6 show, only 20 percent discuss the special care problem. This finding has to be interpreted in the light of the result for Question 7 in Figure 3: More than 60 percent of RCTs were implemented by either the researchers themselves or an NGO. For these cases, a discussion of the special care issue is particularly necessary. The remaining RCTs were implemented by either a large firm or a governmental body – which may better resemble a business-as-usual situation.15
Figure 3: Implementation partners of published RCTs (54 studies included in total, frequencies in bold)

Note: “Regional public authority” refers to interventions implemented by regional governmental entities on the local level.

Table A1 in the online appendix provides a further decomposition of the results presented in Table 1 and shows the share of ‘yes’-answers for the respective year of publication. There is some indication of an improvement from 2009 to 2014, but only for certain questions (Question 1 on people’s awareness of being part in a study and Question 3 on the implications of scaling the share of ‘yes’-answers increases to over 50 percent). For the specific sample dimension the share of ‘yes’ answers to Question 5 is lower in 2014 than in the years from 2009 until 2013. For all other questions, we do not observe major differences. Overall, there is no clear trend towards a systematic and transparent discussion of external validity issues.

5. Discussion

Our interpretation of the results presented in the previous section may be controversial. In this section we consider some of the comments and arguments that have been put forward during the genesis of the paper. We would like to emphasize that for the sake of transparency and rigor we only used objective questions and
abstained from qualitative ratings. While we acknowledge that this approach does not do justice to every single paper, we argue that the overall pattern we obtain is a fair representation of how seriously external validity issues are taken in the publication of RCTs. Also note again that we answered all questions very conservatively.

To summarize the results, we find that many published RCTs do not provide a comprehensive presentation of how the experiment was implemented.\textsuperscript{16} More than half of the papers do not even mention whether the participants in the experiment are aware of being randomized – which is crucial to assess whether Hawthorne- or John-Henry-effects could co-determine the outcomes in the RCT. It is true that in some cases it is obvious whether participants were aware of an experiment or not, but in most cases it is indeed ambiguous. In addition, even in case it is obvious, it is important to know what exactly participants were told and thus, a discussion of how vulnerable the evaluated indicators are to Hawthorne-like distortions would be desirable.

Furthermore, our results show that potential general equilibrium effects are only rarely addressed. This is above all worrisome in case outcomes involve price changes (e.g. labor market outcomes) so that repercussions when the program is brought to scale are almost certain. Likewise, the special care problem is hardly discussed, which is particularly concerning in the developing country context, where many RCTs are implemented by NGOs that are arguably more flexible in terms of treatment provision than the government.

A number of good practice examples exist where external validity issues are avoided by the setting or openly addressed, demonstrating that a transparent discussion of ‘credibility enhancing arguments’ is possible. As for Hawthorne effects, in Karlan et al. (2014) for example participants are not aware of the experiment, which is also clearly stated in the paper. In Bloom et al. (2013), in contrast, participants are aware, but the authors discuss the possibility of distorting effects intensely. For general equilibrium effects, Blattman et al. (2014) address potential adjustments in the equilibrium, which are quite likely in their cash transfer randomization. As for the specific sample problem, Tarozzì et al. (2014) openly discuss that their study might
have taken place in a particular population. Good practice examples for the special care hazard are again BLATTMAN ET AL. (2014), since their program is implemented by the government and therefore resembles a scaled intervention. DUFLO ET AL. (2011a) reveal potential special care problems and acknowledge that a scaled program might be less effective.\textsuperscript{17}

We abstain from giving explicit bad practice examples (for obvious reasons), but indeed some studies are, we believe, negligently silent about certain hazards in spite of very obvious problems. In a minority of cases this is even exacerbated by a very ambitious and broad generalization of the findings.

Some commentators argued that RCTs which test a theory are not necessarily meant to be generalized. Yet, by design we concentrate our review on papers that evaluate a certain policy and hence the vast majority of papers included in this review do generalize results. In addition, a mere test-of-a-theory paper should in our views communicate this clearly to avoid misleading interpretations by policy makers and the public.

This is related to the question whether in fact all papers are supposed to address \textit{all} external validity dimensions included in our review. Our answer is \textit{yes}, at least for policy evaluations that generalize their findings. One might argue that some of the reviewed papers are completely immune to a certain external validity hazard, but the cost of briefly establishing this immunity is negligible (see next Section).

6. Potential Remedies

In an ideal world, external validity would be established by replications in many different populations and using different designs that vary the parameters which potentially codetermine the results. Systematic reviews can then compile the collective information in order to identify patterns in the effectiveness that eventually inform policy. This is the mission of organizations like the Campbell foundation, the Cochrane foundation as well as the International Initiative for Impact Evaluation (3ie) and systematic reviews have indeed been done in a few cases\textsuperscript{18}. In a similar vein, BANERJEE
ET AL. (2016) propose a procedure “from proof of concept to scalable policies”. The authors acknowledge that proof of concept studies are often on purpose conducted under “ideal conditions through finding a context and implementation partner most likely to make the model work”. They suggest an approach of “multiple iterations of experimentation”, in which the context that co-determines the results is refined.

BANERJEE ET AL. (2016) also provide a promising example in India for such a scaling up process. Yet, it is evident that this approach as well as systematic reviews require a massive collective research endeavor that will take many years and that is probably not feasible in all cases.

It is this paper’s stance that in the meantime individual RCTs have to establish external validity, reveal limitations, and discuss implications for transferability openly. To achieve this goal, the first and most obvious step is to include a systematic reporting in RCT-based publications following the CONSORT19 statement in the medical literature. This reference to the CONSORT statement as a role-model for economics research has already been postulated by MIGUEL ET AL. (2014) and EBLE ET AL. (2017), for example. Some design features could be retrieved already in the pre-analysis plan, but at the latest during the peer-review process the checklist should be included and reviewed. Such a checklist ensures that the reader has all information at hand allowing her to make an informed judgement on the transferability of the results. Moreover, potential weaknesses should be disclosed, thereby automatically entailing a qualitative discussion to establish or restrict the study’s external validity. In addition, a mandatory checklist also creates incentives to take external validity issues already into account in the study’s design phase.

Next to more transparence in the publication of RCTs, a few instruments exist to deal with external validity hazards – some of which are post hoc, some can be incorporated in the design of the study. For Hawthorne and John Henry effects, the most obvious solution is not to inform the participants about the randomization, which of course hinges upon the study design. Such an approach resembles to what LEVITT AND LIST (2009) refer to as a ‘natural field experiment’. In some set-ups people have to be informed, either because randomization is obvious or for ethical reasons. The standard
remedy in medical research – assigning a third group to a placebo treatment – is not possible in most experiments in social sciences. ALDASHEV et al. (2017) emphasize that the assignment procedure that is used to randomly assign participants into treatment and control groups affects the size of the bias considerably. They suggest that a public randomization reduces bias as compared to a non-transparent private randomization. Accounting for general equilibrium effects comprehensively is impossible in most cases, since all types of macro-economic adjustments can hardly be captured in a micro-economic study. In order to evaluate what eventually happens in the general equilibrium one would have to resort to computable general equilibrium (CGE) models. Indeed, there are ambitions to plug the results of RCT based evaluations into CGE models as it is done with observational data in COADY AND HARRIS (2004).

The seminal work on GEE so far tests for the existence of at least selected macro-economic adjustments and spillovers by randomizing not only the treatment within clusters (e.g. markets), but also the treatment density between clusters. Influential examples for this approach are CRÉPON ET AL. (2013) on the French labor market and MURALIDHARAN AND SUNDARARAMAN (2015) for school vouchers in India. Using the same approach, BURKE (2014) randomizes the density of loan offers across regions to account for GEE. Moreover, randomizing the intervention on a higher regional aggregation allows for examining the full general equilibrium effect at that level (BANERJEE ET AL. 2016). MURALIDHARAN ET AL. (2017), for example, examine a public employment program at a regional level that is “large enough to capture general equilibrium effects”. ATTANASIO ET AL. (2011) exploit the randomized PROGRESA rollout on the village level to study GEE on child wages.

As for the specific sample problem, there is an emerging literature that provides guidance on extrapolating findings from one region to another. PEARL AND BAREINBOIM (2014) develop a conceptual framework that enables the researcher to decide whether transferring results between populations is possible at all. Moreover, they formulate assumptions that, if they hold true, allow for transferring results from RCT based studies to observational ones (“license to transport”). GECHTER (2016) takes a similar line and develops a methodology that calculates bounds for transferring treatment
effects obtained in an RCT to a non-experimental sample. The key assumption here is that “the distribution of treated outcomes for a given untreated outcome in the context of interest is consistent with the experimental results” (see Gechter 2016:3). Further contributions offer solutions for very specific types of RCTs. Kowalski (2016) provides a methodology suitable for RCTs using an encouragement design (this is, with low compliance rates), Stuart et al. (2011) propose a methodology to account for selection into the RCT sample, which is often the case in effectiveness studies.

The degree to which scholars believe in the generalizability of results also hinges upon which part of the results chain they focus on. One line of thinking concentrates on the human behavior component in evaluations, also referred to as ‘mechanism’, and assumes this to be more generalizable than what is found on the intervention as a whole (see for example Bates and Glennister 2017). The other viewpoint puts more emphasis on the treatment as a policy intervention. Here, the complexity of interventions and the context in which they happen are decisive. This camp calls for combining evidence from rigorous evaluations with case studies (Woolcock 2013) or ‘reasoned intuition’ (Basu 2014; Basu and Foster 2015) to transfer findings from one setting to a different policy population.

This complexity feature is very much related to what we have referred to as special care in the provision of the treatment, which is arguably very heterogeneous across different policy environments. There seems to be a growing consensus that this is an important external validity concern (see for example Banerjee et al. 2016) and some scholars have made recommendations on how to account for this. Both Bates and Glennister (2017) and Woolcock (2013) provide frameworks that guide the transferability assessment and special care is one important feature. Bates and Glennister (2017) suggest isolating the mechanism from other intervention related features, while Woolcock (2013) argues that in many “developing countries […] implementation capability is demonstrably low for logistical tasks, let alone for complex ones”. Hence, the higher the complexity of an intervention, the more implementation capability becomes a bottleneck and, to use our wording, the more special care puts external validity at risk. Woolcock’s position is that for complex
interventions – so the vast majority of policy interventions – generalizing is a “decidedly high-uncertainty undertaking”. He suggests including qualitative case studies into these deliberations.

7. Conclusion

In theory, there seems to be a consensus among empirical researchers that establishing external validity of a policy evaluation is as important as establishing its internal validity. Against this background, this paper has systematically reviewed published RCTs to examine whether external validity concerns are addressed. Our findings suggest that external validity is often neglected and does not play the important role that it is associated with in review papers and the general academic debate.

In a nutshell, our sole claim is that papers should discuss the extent to which the different hazards to external validity apply. We call for dedicating the same devotion to establishing external validity as is done to establishing internal validity. This thinking implies that papers published in top academic journals are not only targeted to the research community, but also to a policy-oriented audience (including decision-makers and journalists). This audience, in particular, needs all the information necessary to make informed judgements on the extent to which the findings are transferable to other regions and non-experimental business-as-usual settings. A more transparent reporting would also lead to a situation in which more generalizable RCTs receive more attention than those that were implemented under heavily controlled circumstances or in a very specific region only.

It would be desirable if the peer review process at economics journals explicitly scrutinized design features of RCTs that are relevant for generalization. As a starting point, this does not need to be more than a checklist and short statements to be included in an electronic appendix. The logic is that only if researchers know already at the beginning of a study that they will need to provide such checklists and discussions, they will have clear incentives to account for external validity issues in the study design. Otherwise, external validity degenerates to a nice-to-have feature
that researchers account for voluntarily and for intrinsic reasons. These internal incentives will probably work in many cases. But given the trade-offs we all face during the laborious implementation of studies it is almost certain that external validity will often be sacrificed for other features to which the peer-review process currently pays more attention.
Endnotes

1 Our paper’s title is an obvious reminiscence to an important contribution to this debate, Angus Deaton’s “Instruments of Development: Randomization in the Tropics and the Search for the Elusive Keys to Economic Development”, published as an NBER working paper in 2009. A revised version was published under a different title in the Journal of Economic Literature (DEATON, 2010).

2 Note that our focus is on policy evaluation. In our protocol, we therefore excluded laboratory experiments, framed field experiments, and test-of-a-theory field experiments that are obviously not meant to evaluate a policy intervention.

3 The Hawthorne effect in some cases cannot be distinguished from survey effects, the Pygmalion effect, and the observer-expectancy effect (see BULTE ET AL. 2014). All of these effects, which generally also might occur in observational studies, can be amplified by the Hawthorne effect and the experimental character of the study. See ALDASHEV ET AL. (2017) for a formalization of the Hawthorne and John Henry effect.

4 The John Henry effect describes the effect that being randomized into the control group can have on the performance of control group members. John Henry is a legendary black railroad worker, who – equipped with a traditional hammer – competed with a steam drill in an experimental setting. Being aware of this exercise he strived to outperform the steam drill. While he eventually succeeded, he died from exhaustion (see SARETSKY 1972, also for a very classic example of a John Henry effect).


6 See CRÉPON ET AL. (2013) for an example of such GEE in a randomized labor market program, in which treated participants benefited at the expense of non-treated participants.

7 ATTANASIO ET AL. (2011) observe a reduction in labor supply for child labor in the Mexican PROGRESA conditional cash transfer intervention, which is disbursed conditioned on children going to school.

8 The present study builds on an earlier paper that also included RCTs conducted in developed countries, see PETERS, LANGBEIN, AND ROBERTS (2016).

9 See Appendix B for the list of the excluded papers and the reason for exclusion.

10 A comprehensive list of included papers and their rating is found in Appendix A.

11 See ROETMAN (2011) for more information on the genesis of RCTs in Kenya and the role of ICS.

12 The filter question on whether the paper generalizes beyond the study population was added post-hoc, as a response to comments made by some authors.

13 The time period of a study is of course not only an external validity issue. See KING AND BEHRMAN (2009) on the relevance of timing for impact evaluations.

14 We coded this question by ‘yes’ in case the paper derives explicit policy recommendations for other regions or countries and in case it makes statements like “our results suggest that this policy works/does not work” or “our results generalize to”.

15 It could of course be argued that NGOs can also be considered as “business-as-usual”, since many real-world interventions, especially in developing countries, are implemented by NGOs. However, for most of the 20 RCTs that were implemented by an NGO, the cooperating NGO was a rather small one and regionally limited in its activities. Thus, bringing the intervention to scale would be the task of either the government or a larger NGO with potential implications for the efficacy of the intervention.

16 This finding is in line with Eble et al. (2017) who review RCTs published between 2001 and 2011 for how they deal with different sorts of biases (also covering Hawthorne effects).

17 Details on these examples can be found in the review report on the respective paper in the Online Appendix.


## Appendix

### Appendix A: Reviewed papers and ratings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aker et al. (2012)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>NGO</td>
<td>Yes</td>
</tr>
<tr>
<td>Alatas et al. (2012)</td>
<td>Yes</td>
<td>Participants are NOT aware</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Government</td>
<td>No</td>
</tr>
<tr>
<td>Ashraf et al. (2010)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>NGO</td>
<td>No</td>
</tr>
<tr>
<td>Ashraf et al. (2014)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>No</td>
</tr>
<tr>
<td>Attanasio et al. (2011)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Government</td>
<td>Yes</td>
</tr>
<tr>
<td>Baird et al. (2011)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>NGO</td>
<td>No</td>
</tr>
<tr>
<td>Barrera-Osorio et al. (2011)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Regional Public Authority</td>
<td>Yes</td>
</tr>
<tr>
<td>Bertrand et al. (2010)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Firm</td>
<td>No</td>
</tr>
<tr>
<td>Björkman, Svennsson (2009)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>NGO</td>
<td>Yes</td>
</tr>
<tr>
<td>Blattman et al. (2014)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Government</td>
<td>Yes</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------------------------------</td>
<td>-------------------------------</td>
<td>------------------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>-------------------------------------------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Blimpo (2014)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Researcher</td>
</tr>
<tr>
<td>Bloom et al. (2013)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Researcher</td>
</tr>
<tr>
<td>Burde, Linden (2013)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>NGO</td>
<td>No</td>
</tr>
<tr>
<td>Casey et al. (2012)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Government</td>
</tr>
<tr>
<td>Chinkhumba et al. (2014)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Cohen, Dupas (2010)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Collier, Vicente (2014)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>NGO</td>
<td>Yes</td>
</tr>
<tr>
<td>Das et al. (2013)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NGO</td>
<td>Yes</td>
</tr>
<tr>
<td>de Mel et al. (2009a)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>No</td>
</tr>
<tr>
<td>de Mel et al. (2013)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>No</td>
</tr>
<tr>
<td>Drexler et al. (2014)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Firm</td>
<td>No</td>
</tr>
<tr>
<td>Dufo et al. (2011a)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NGO</td>
<td>Yes</td>
</tr>
</tbody>
</table>

28
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Duflo et al. (2011b)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>NGO</td>
</tr>
<tr>
<td>Duflo et al. (2012)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NGO</td>
</tr>
<tr>
<td>Duflo et al. (2013)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Regional Public Authority</td>
<td>Yes</td>
</tr>
<tr>
<td>Dupas (2011)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NGO</td>
<td>Yes</td>
</tr>
<tr>
<td>Dupas, Robinson (2013a)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Firm</td>
<td>Yes</td>
</tr>
<tr>
<td>Dupas, Robinson (2013b)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Dupas (2014)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Feigenberg et al. (2013)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Firm</td>
<td>Yes</td>
</tr>
<tr>
<td>Field et al. (2013)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Firm</td>
<td>Yes</td>
</tr>
<tr>
<td>Fujiwara, Wantchekon (2013)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Giné et al. (2010)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Firm</td>
<td>Yes</td>
</tr>
<tr>
<td>Giné et al. (2012)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Government</td>
<td>Yes</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------</td>
<td>-------------------------------</td>
<td>----------------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>------------------------------------------------</td>
<td>------------------------------------</td>
<td>----------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Glewwe et al. (2009)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Glewwe et al. (2010)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Hanna et al. (2014)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Jensen (2010)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Jensen (2012)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Jensen and Miller (2011)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Government</td>
<td>Yes</td>
</tr>
<tr>
<td>Karlan et al. (2014)</td>
<td>Yes</td>
<td>Participants are NOT aware</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Government</td>
<td>No</td>
</tr>
<tr>
<td>Karlan and Valdivia (2011)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>NGO</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Kremer et al. (2011)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>NGO</td>
<td>No</td>
</tr>
<tr>
<td>Kremer et al. (2009)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NGO</td>
<td>Yes</td>
</tr>
<tr>
<td>Macours et al. (2012)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Government</td>
<td>No</td>
</tr>
<tr>
<td>Macours, Renos (2014)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Government</td>
<td>No</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------------------------------------------</td>
<td>-------------------------------</td>
<td>------------------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td>--------------------------------</td>
<td>-----------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Muralidharan, Venkatesh (2010)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>NGO</td>
<td>No</td>
</tr>
<tr>
<td>Muralidharan, Venkatesh (2011)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NGO</td>
<td>No</td>
</tr>
<tr>
<td>Olken et al. (2014)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Government</td>
<td>Yes</td>
</tr>
<tr>
<td>Oster, Thornton (2011)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Pradhan et al. (2014)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Firm</td>
<td>Yes</td>
</tr>
<tr>
<td>Robinson (2012)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Researcher</td>
<td>Yes</td>
</tr>
<tr>
<td>Tarozzi et al. (2014)</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Firm</td>
<td>Yes</td>
</tr>
<tr>
<td>Vicente (2014)</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Government</td>
<td>Yes</td>
</tr>
</tbody>
</table>
### Appendix B: Excluded papers and reason for exclusion

<table>
<thead>
<tr>
<th>Author</th>
<th>Reason for exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adhvaryu (2014)</td>
<td>Quasi-Experiment</td>
</tr>
<tr>
<td>Armantier and Boly (2013)</td>
<td>Artefactual Experiment</td>
</tr>
<tr>
<td>Ashraf (2009)</td>
<td>Behavioral Field Experiment</td>
</tr>
<tr>
<td>Attanasio et al. (2012)</td>
<td>Artefactual Experiment</td>
</tr>
<tr>
<td>Bauer et al. (2012)</td>
<td>Behavioral Field Experiment</td>
</tr>
<tr>
<td>Beaman and Magruder (2012)</td>
<td>Artefactual Experiment</td>
</tr>
<tr>
<td>Beaman et al. (2010)</td>
<td>Natural Experiment</td>
</tr>
<tr>
<td>Besley et al. (2012)</td>
<td>Theoretical Paper</td>
</tr>
<tr>
<td>Bursztyn and Coffman (2012)</td>
<td>Natural Experiment</td>
</tr>
<tr>
<td>Cai et al. (2009)</td>
<td>Behavioral Field Experiment</td>
</tr>
<tr>
<td>Chassang et al. (2012)</td>
<td>Theoretical paper about RCTs</td>
</tr>
<tr>
<td>De Mel et al. (2009b)</td>
<td>Reply to a previously published article</td>
</tr>
<tr>
<td>Di Tella and Schargrodsky</td>
<td>Natural Experiment</td>
</tr>
<tr>
<td>Gneezy et al. (2009)</td>
<td>Artefactual Experiment</td>
</tr>
<tr>
<td>Hjort (2014)</td>
<td>Natural Experiment</td>
</tr>
<tr>
<td>Karlan et al. (2011)</td>
<td>Behavioral Field Experiment</td>
</tr>
<tr>
<td>Karlan and Zinman (2009)</td>
<td>Behavioral Field Experiment</td>
</tr>
<tr>
<td>Lucas, Mbiti (2014)</td>
<td>Quasi-Experiment</td>
</tr>
<tr>
<td>Voors et al. (2012)</td>
<td>Artefactual Experiment</td>
</tr>
</tbody>
</table>
References


Cartwright, N. 2010. 'What are Randomised Controlled Trials Good for?', *Philosophical Studies*, 147: 59-70.


