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

Jörg Peters, Jörg Langbein, and Gareth Roberts

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), we scrutinize the following hazards to external validity: 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 codes: C83, C93

In recent years, intense debate has taken place about the value of Randomized Controlled Trials (RCTs).[1] Most notably in development economics, RCTs have assumed a dominant 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 (Moffit 2004; Roe and Just 2009; and Temple 2010; Dehejia 2015; Muller 2015; Prittchet and Sandefur 2015). 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

The World Bank Research Observer

doi: 10.1093/wbro/lkx005 33:34–64

resource-constrained governments, while the original RCT is often implemented by effective NGOs or the researchers themselves (Ravallion 2012; Bold et al. 2013; Banerjee et al. 2017; Deaton and Cartwright 2016).

This does not question the enormous contribution that RCTs have made to existing knowledge about the effectiveness of policy interventions. Rather, it 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 that appeared 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: (a) Hawthorne and John Henry Effects; (b) general equilibrium effects; (c) specific sample problems; and (d) 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 scrutinized 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 with an eye toward whether they address 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). As early as 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). Indeed, Banerjee and Rodrik seem to agree that external validity is never a self-evident fact in empirical research, and that RCTs in particular should 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. Following that, the methodological approach and the seven questions are discussed. The results are presented in the next section, followed by a discussion section. The subsequent section provides an overview on existing remedies for external validity problems and ways to deal with them in practice. The final section concludes.

Theoretical Background and Definition of External Validity

Theoretical Framework

Understanding 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) defines external validity in a way that is similar to the understanding conveyed in Duflo, Glennerster, and Kremer (2008): "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, and $\bar{C}$ denotes the absence of the intervention. 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: (a) A needs to be a representative sample of A'; (b) C is introduced in A' as it was in the experiment in A; (c) 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 Duflo, Glennerster, and Kremer (2008). Similar to the Cartwright framework, Duflo, Glennerster, and Kremer 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". The four hazards to external validity that are identified by Duflo, Glennerster, and Kremer 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 in every case, 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.

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 Duflo, Glennerster, and Kremers' 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 are under observation.[3] It is obvious that this could lead to 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 setup, distortions of participants'

behavior are less likely. In contrast, if the randomized intervention interferes noticeably with the participants' daily life (e.g., 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, Kremer, and Robinson (2011b). The major difference is that the program examined in Bold et al. was implemented by the national government instead of an NGO, as was the case in the Duflo et al. study. The positive results observed in Duflo, Kremer, and Robinson (2011b) could not be replicated in Bold et al. (2013): "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."

Vivalt (2017) 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), who 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 refers to as "site selection bias", whereby 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 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 initially cooperate with researchers.

The third identified hazard arises from potential general equilibrium effects (GEE).[6] 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.[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 that the study examines. 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, 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 being equal, the generalizability of the results from observational studies is also exposed by potential GEE. Many RCTs, particularly 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 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.

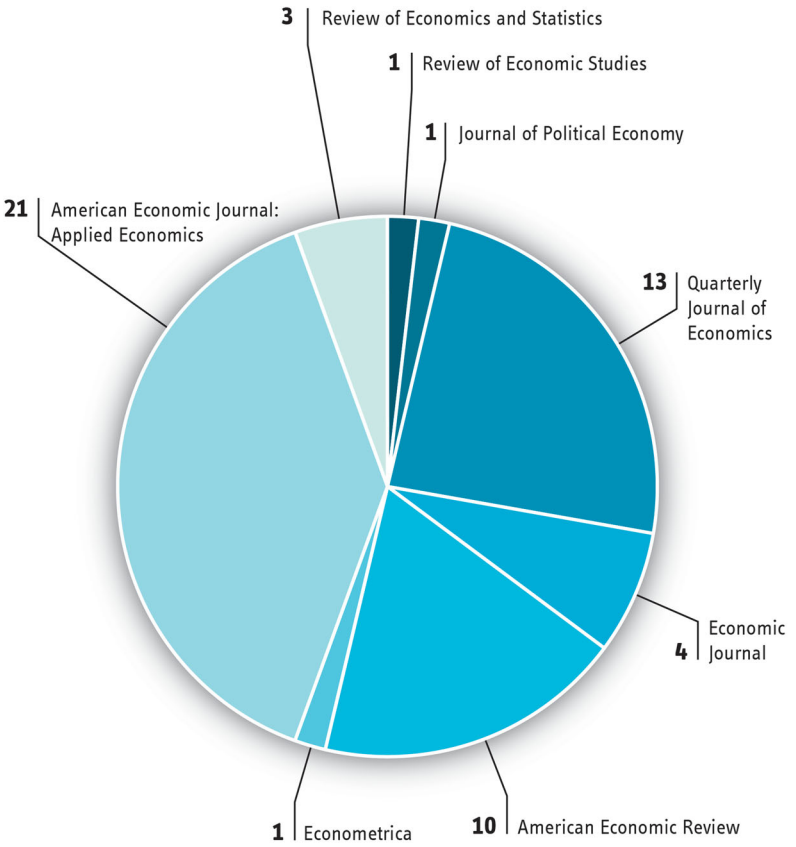
Methods and Data

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*, *Quarterly Journal of Economics*, *Journal of Political Economy,* the *Review of Economic Studies,* as well as further leading journals that publish empirical work using RCTs such as *American Economic Journal: Applied Economics, Economic Journal*, and *Review of Economics and Statistics*.

We scrutinized all issues in the period, particularly all papers that mention either the terms "field experiment", "randomized controlled trials", or "experimental evidence" in either the title or the abstract, or which indicated in the abstract or the title

Figure 1. Published RCTs Between 2009 and 2014

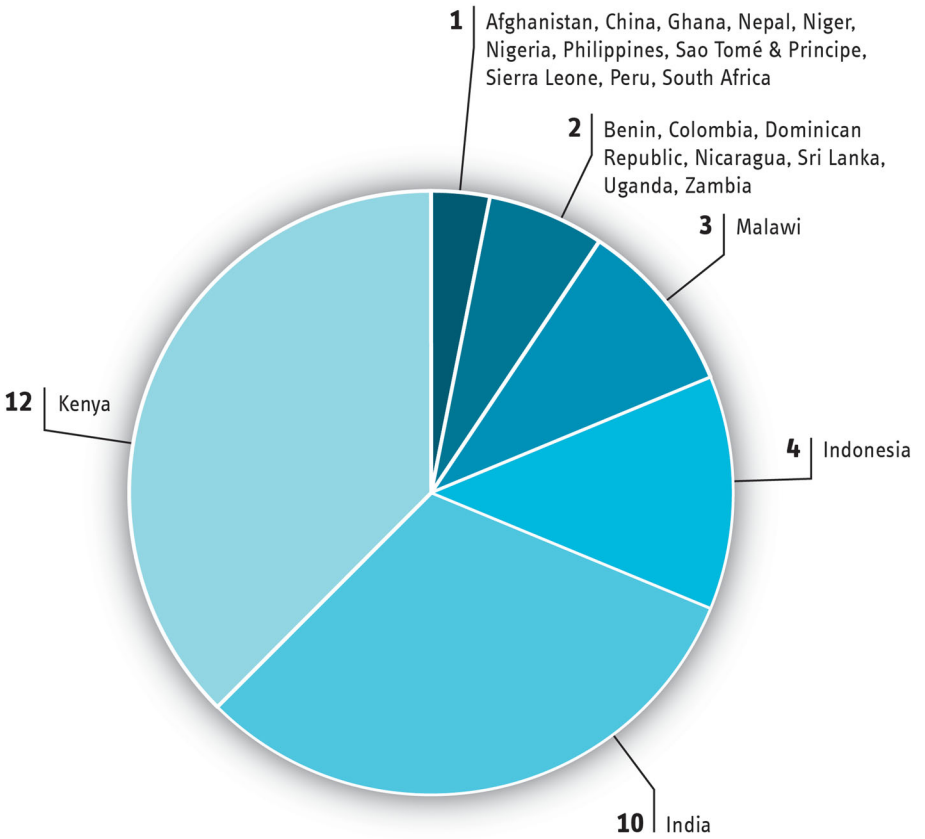


Note: A total of 54 studies were included, frequencies appear in bold.

that a policy intervention was randomly introduced. We excluded those papers that examine interventions in an OECD member country.[8] In total, 73 papers were initially identified. Our focus is on policy evaluation and we therefore excluded mere test-of-a-theory papers.[9] 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.[10] The distribution across journals is uneven, with the vast majority being published in *American Economic Journal: Applied Economics*, *American Economic Review* and *Quarterly Journal of Economics* (see figure 1).

Figure 2 depicts the regional coverage of the surveyed RCTs. The high 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

Figure 2. Countries of Implementation



Note: A total of 54 studies were included, frequencies appear in bold.

for Poverty Action [IPA] and the Abdul Latif Jameel Poverty Action Lab [J-Pal]). Most of these 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.[11]

We read all 54 papers carefully (including the online supplementary appendix) to determine whether each paper addressed 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 we intentionally 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 the first, and examine whether the

"credibility-enhancing arguments" (Rodrik 2009) are provided to underpin the plausibility of external validity. Appendix A in the appendix shows the answers to the seven questions for all surveyed papers individually. In general, we answered the questions conservatively, that is, 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).[12] 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 supplementary appendix to this paper.

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 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 do we additionally pose the following 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 (in the design of the study, in the interpretation of the treatment/mechanisms, or in the interpretation of the size of the impact)?

The next set of questions probes into *general equilibrium effects*. As outlined in the first section, 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 through 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 Duflo, Glennerster, and Kremer 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 the filter question below). 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 judgment on whether we deem the statement to be plausible and the discussion to be comprehensive.
The fourth hazard, special care, is accounted for by the last 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 judgment 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 ask:

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.[14] Note that in this question we do not assess the

Table 1. Reporting on External Validity in Published RCTs

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

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

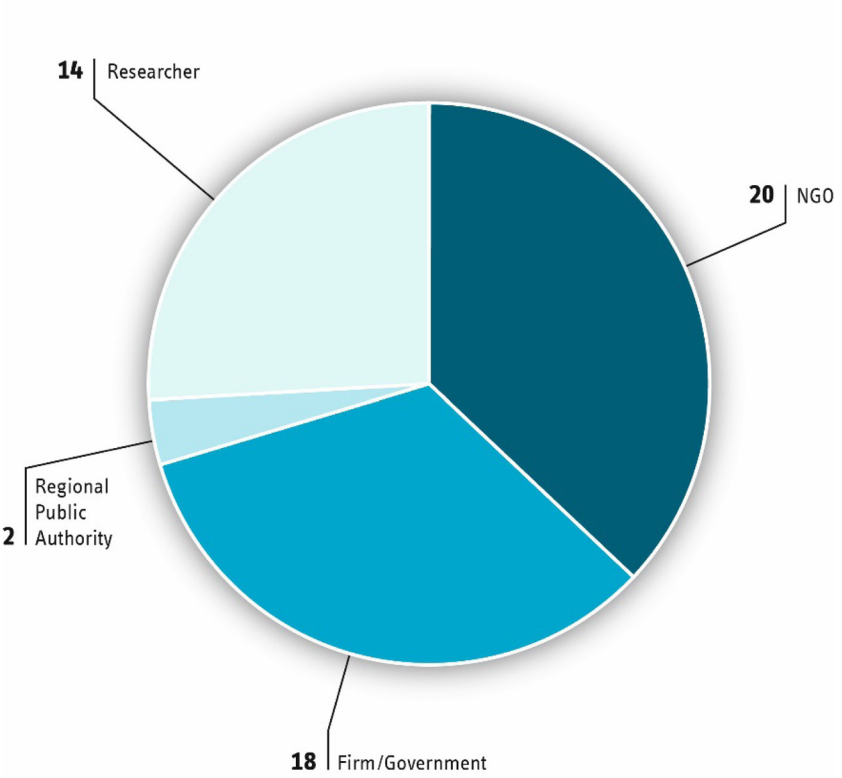
degree to which the paper generalizes (which in fact varies considerably), but only if it generalizes *at all*.

Results

Table 1 shows the results for the seven questions asked of 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 either 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

Figure 3. Implementation Partners of Published RCTs



Note: "Regional public authority" refers to interventions implemented by regional governmental entities on the local level. A total of 54 studies were included, frequencies appear in bold.

effects are those that in fact examine 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, that is, received a "no" for questions 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.

As the results for question 6 show, only 20 percent discuss the special care problem. This finding has to be interpreted in 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 relevant. The

remaining RCTs were implemented by either a large firm or a governmental body—which may better resemble a business-as-usual situation.[15]

Table A1 in the supplementary 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. For example, the share of "yes"-answers increases to over 50 percent for question 1 on people's awareness of being part in a study and question 3 on the implications of scaling. For the specific sample dimension, the share of "yes" answers to question 5 is lower in 2014 than in it is 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.

Discussion

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. Please note once 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.[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 for assessing whether Hawthorne or John Henry effects could co-determine the outcomes in the RCT. It is true that in some cases it is obvious that participants were aware of an experiment, but in most cases it is indeed ambiguous. In addition, even in cases where 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 the case that 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, Fiala, and Martinez (2014) address potential adjustments in the equilibrium, which are quite likely in their cash transfer randomization. As for the specific sample problem, Tarozzi 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, Fiala, and Martinez (2014), since their program is implemented by the government and therefore resembles a scaled intervention. Duflo, Dupas, and Kremer (2011a) reveal potential special care problems and acknowledge that a scaled program might be less effective.[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 that 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 of whether in fact *all* papers are supposed to address *all* external validity dimensions included in our review. Our answer is *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.

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.[18] In a similar vein, Banerjee et al. (2017) propose a procedure "from proof of concept to scalable policies." The authors acknowledge that proof of concept studies are often intentionally conducted under "ideal conditions through finding a context and implementation partner most likely to make the model work". These authors suggest an approach of "multiple iterations of experimentation", in which the context that

co-determines the results is refined. Banerjee et al. (2017) 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 is probably not feasible in all cases.

It is this paper's stance that in the meantime, individual RCTs with a claim to broader policy relevance 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 CONSORT statement in the medical literature.[19] This reference to the CONSORT statement as a role-model for economics research has already been postulated by Miguel et al. (2014) and Eble, Boone, and Elbourne (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 judgment 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 already take external validity issues 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, others of which 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 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, Kirchsteiger, and Sebald (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. These authors suggest that a public randomization reduces bias 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 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 macroeconomic adjustments and spillovers by randomizing not only the treatment within clusters (e.g., markets), but also the treatment density between clusters. Influential

examples of 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 et al. (2017) 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. 2017). Muralidharan, Niehaus, and Sukh (2017), for example, examine a public employment program at a regional level that is "large enough to capture general equilibrium effects". Attanasio, Kugler, and Meghir (2011) exploit the randomized PROGRESA roll-out on the village level to study GEE on child wages.

As for the specific sample problem, there is an emerging body of 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, these authors 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). Further contributions offer solutions for very specific types of RCTs. For example, Kowalski (2016) provides a methodology suitable for RCTs using an encouragement design (i.e., with low compliance rates), while 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. 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, e.g., Bates and Glennerster 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, e.g., Banerjee et al. 2017), and some scholars have made recommendations on how to account for this. Both Bates and Glennerster (2017) and Woolcock (2013) provide frameworks that guide the transferability assessment, and special care is one important feature.

Bates and Glennerster (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—that is, the vast majority of policy interventions—generalizing is a "decidedly high-uncertainty undertaking". Woolcock suggests including qualitative case studies into these deliberations.

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 when 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, requires all the information necessary to make informed judgments on the extent to which the findings are transferable to other regions and non-experimental business-as-usual settings. 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 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.

Notes

Jörg Peters is heading the research group "Climate Change in Developing Countries" at RWI, Germany and is Professor at University of Passau. All correspondence to be sent to: Jörg Peters, RWI, Hohenzollernstraße 1–3, 45128 Essen, Germany, e-mail: peters@rwi-essen.de, phone: 49-201-8149-247. Jörg Langbein is *Researcher* at RWI, Germany. Gareth Roberts is lecturer at University of the Witwatersrand and researcher at AMERU, Johannesburg, South Africa. The authors thank Maximilian Huppertz and Julian Rose for excellent research assistance. The authors are also grateful for valuable comments and suggestions by the editor Peter Lanjouw, three anonymous referees, Martin Abel, Mark Andor, Michael Grimm, Angus Deaton, Heather Lanthorn, Luciane Lenz, Stephan Klasen, Laura Poswell, and Colin Vance, as well as seminar participants at the University of Göttingen, University of Passau, University of the Witwatersrand, and Stockholm Institute of Transition Economics. We contacted all lead authors of the papers included in this review and many of them provided helpful comments on our manuscript. Langbein and Peters gratefully acknowledge the support of a special grant (Sondertatbestand) from the German Federal Ministry for Economic Affairs and Energy and the Ministry of Innovation, Science, and Research of the State of North Rhine-Westphalia. A supplemental appendix to this article is available at https://academic.oup.com/WBRO.

1. The title is an obvious reference 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 (Deaton 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, Kirchsteiger, and Sebald (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 sledgehammer—competed with a steam trill 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, for a very classic example of a John Henry effect).
5. See Bulte et al. (2014) and Simons et al. (2017) for evidence on strong Hawthorne effects in experiments in Tanzania and Uganda, respectively, and McCambridge, Witton, and Elbourne (2014) for a systematic review on Hawthorne effects in medical research. Cilliers, Dube, and Siddiqi (2015) provide evidence for the distorting effects of foreigner presence in framed field experiments in developing countries. See also Zwane et al. (2011).
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, Kugler, and Meghir (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, Boone, and Elbourne (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 supplementary appendix.

18. Examples of systematic reviews are Acharya et al. (2012) on health insurance for the informal sector, Evans and Popova (2016) on school learning, Evans and Popova (2017) on cash transfers, and McKenzie and Woodruff (2013) on the impacts of business training interventions. See also the 3ie systematic review data base available at: www.3ieimpact.org/en/evidence/systematic-reviews/.

19. See Moher et al. (2010) and Schulz, Altman, and Moher (2010).

References

Acharya, A., S. Vellakkal, F. Taylor, E. Masset, A. Satija, M. Burke, and S. Ebrahim. 2012. "The Impact of Health Insurance Schemes for the Informal Sector in Low- and Middle-Income Countries: A Systematic Review." *World Bank Research Observer* 28 (2): 236–66.

Adhvaryu, A. 2014. "Learning, Misallocation, and Technology Adoption." *Review of Economic Studies* 81 (4): 1331–65.

Aker, J. C., C. Ksoll, and T. J. Lybbert. 2012. "Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger." *American Economic Journal: Applied Economics* 4 (4): 94–120.

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review* 102 (4): 1206–40.

Aldashev, G., G. Kirchsteiger, and A. Sebald. 2017. "Assignment Procedure Biases in Randomised Policy Experiments." *The Economic Journal* 127 (602): 873–95.

Allcott, H. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117–65.

Armantier, O., and A. Boly. 2013. "Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada." *Economic Journal* 123 (573): 1168–87.

Ashraf, N. 2009. "Spousal Control and Intra-household Decision Making: An Experimental Study in the Philippines." *American Economic Review* 99 (4): 1245–77.

Ashraf, N., J. Berry, and J. M. Shapiro. 2010. "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia." *American Economic Review* 100 (5): 2383–413.

Ashraf, N., E. Field, and J. Lee. 2014. "Household Bargaining and Excess Fertility: An Experimental Study in Zambia." *American Economic Review* 104 (7): 2210–37.

Attanasio, O., A. Kugler, and C. Meghir. 2011. "Subsidizing Vocational Training for Disadvantaged Youth in Colombia: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics* 3 (3): 188–220.

Attanasio, O., A. Barr, J. C. Cardenas, G. Genicot, and C. Meghir. 2012. "Risk Pooling, Risk Preferences, and Social Networks." *American Economic Journal: Applied Economics* 4 (2): 134–67.

Banerjee, A. V. 2005. "New Development Economics and the Challenge to Theory." In *New Directions in Development Economics: Theory or Empirics? A Symposium in Economic and Political Weekly*, edited by R. Kanbur, unpublished paper, August 2005.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukherji, M. Shotland, and M. Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives* 31 (4): 73–102.

Baird, S., C. McIntosh, and B. Özler. 2011. "Cash or Condition? Evidence from a Cash Transfer Experiment." *Quarterly Journal of Economics* 126 (4): 1709–53.

Barrera-Osorio, F., M. Bertrand, L. L. Linden, and F. Perez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia." *American Economic Journal: Applied Economics* 3 (2): 167–95.

Basu, K. 2014. "Randomisation, Causality and the Role of Reasoned Intuition." *Oxford Development Studies* 42 (4): 455–72.

Basu, K., and A. Foster. 2015. "Development Economics and Method: A Quarter Century of ABCDE." *World Bank Economic Review* 29 (suppl_1): S2–S8.

Bates, M. A., and R. Glennerster. 2017. "The Generalizability Puzzle." *Stanford Social Innovation Review*, Summer 2017: 50–54.

Bauer, M., J. Chytilová, and J. Morduch. 2012. "Behavioral Foundations of Microcredit: Experimental and Survey Evidence from Rural India." *American Economic Review* 102 (2): 1118–39.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics* 124 (4): 1497–540.

Beaman, L., and J. Magruder. 2012. "Who Gets the Job Referral? Evidence from a Social Networks Experiment." *American Economic Review* 102 (7): 3574–93.

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *Quarterly Journal of Economics* 125 (1): 263–306.

Besley, T. J., K. B. Burchardi, and M. Ghatak. 2012. "Incentives and the De Soto Effect." *Quarterly Journal of Economics* 127 (1): 237–82.

Björkman, M., and J. Svensson. 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda." *Quarterly Journal of Economics* 124 (2): 735–59.

Blattman, C., N. Fiala, and S. Martinez. 2014. "Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda." *Quarterly Journal of Economics* 129 (2): 697–752.

Blimpo, M. P. 2014. "Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin." *American Economic Journal: Applied Economics* 6 (4): 90–109.

Bloom, N., B. Eifer, A. Mahajan, D. McKenzie, and J. Roberts. 2013. "Does Management Matter? Evidence from India." *Quarterly Journal of Economics* 128 (1): 1–51.

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." Working Paper No. 321, Center for Global Development, Washington, DC.

Bulte, E., G. Beekman, S. di Falco, J. Hella, and P. Lei. 2014. "Behavioral Responses and the Impact of new Agricultural Technologies: Evidence from a Double-Blind Field Experiment in Tanzania." *American Journal of Agricultural Economics* 96 (3): 813–30.

Burde, D., and L. L. Linden. 2013. "Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools." *American Economic Journal: Applied Economics* 5 (3): 27–40.

Burke, M., L. F. Bergquist, and E. Miguel. 2017. "Selling Low and Buying High: An Arbitrage Puzzle in Kenyan Villages." Working Paper, UC Berkeley. Accessed January 17, 2018. Available at: https://web.stanford.edu/~mburke/papers/MaizeStorage.pdf.

Bursztyn, L., and L. C. Coffman. 2012. "The Schooling Decision: Family Preferences, Intergenerational Conflict, and Moral Hazard in the Brazilian Favelas." *Journal of Political Economy* 120 (3): 359–97.

Cai, H., Y. Chen, and H. Fang. 2009. "Observational Learning: Evidence from a Randomized Natural Field Experiment." *American Economic Review* 99 (3): 864–82.

Cartwright, N. 2010. "What are Randomised Controlled Trials Good For?" *Philosophical Studies* 147: 59–70.

Casey, K., R. Glennerster, and E. Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan." *Quarterly Journal of Economics* 127 (4): 1755–812.

Chassang, S., I. Miquel, G. P., and E. Snowberg. 2012. "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments." *American Economic Review* 102 (4): 1279–309.

Chinkhumba, J., S. Godlonton, and R. Thornton. 2014. "The Demand for Medical Male Circumcision." *American Economic Journal: Applied Economics* 6 (2): 152–77.

Cilliers, J., O. Dube, and B. Siddiqi. 2015. "The White-Men Effect: How Foreigner Presence Affects Behavior in Experiments." *Journal of Economic Behavior and Organization* 118: 397–414.

Coady, D. P., and R. L. Harris. 2004. "Evaluating Transfer Programmes within a General Equilibrium Framework." *Economic Journal* 114 (498): 778–99.

Cohen, J., and P. Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125 (1): 1–45.

Collier, P., and P. C. Vicente. 2014. "Votes and Violence: Evidence from a Field Experiment in Nigeria." *Economic Journal* 124 (574): F327–55.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora. 2013. "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128 (2): 531–80.

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman. 2013. "School Inputs, Household Substitution, and Test Scores." *American Economic Journal: Applied Economics* 5 (2): 29–57.

de Mel, S., D. McKenzie, and C. Woodruff. 2009a. "Returns to Capital in Microenterprises: Evidence from a Field Experiment." *Quarterly Journal of Economics* 124 (1): 423.

———. 2009b. "Are Women More Credit Constrained? Experimental Evidence on Gender and Microenterprise Returns." *American Economic Journal: Applied Economics* 1 (3): 1–32.

———. 2013. "The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka." *American Economic Journal: Applied Economics* 5 (2): 122–50.

Deaton, A. S. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." NBER Working Paper No. 14690, National Bureau of Economic Research, Cambridge, MA.

——— 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

Deaton, A. S., and N. Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." NBER Working Paper No. 22595, National Bureau of Economic Research, Cambridge, MA.

Deheija, R. 2015. "Experimental and Non-Experimental Methods in Development Economics: A Porous Dialectic." *Journal of Globalization and Development* 6 (1): 47–69.

DiTella, R., and E. Schargrodsky. 2013. "Criminal Recidivism after Prison and Electronic Monitoring." *Journal of Political Economy* 121 (1): 28–73.

Drexler, A., G. Fischer, and A. Schoar. 2014. "Keeping It Simple: Financial Literacy and Rules of Thumb." *American Economic Journal: Applied Economics* 6 (2): 1–31.

Duflo, E., P. Dupas, and M. Kremer. 2011a. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5): 1739–74.

Duflo, E., R. Glennerster, and M. Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, edited by P. Schultz and J. Strauss, 3895–962, Amsterdam: North Holland.

Duflo, E., M. Greenstone, R. Pande, and N. Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *Quarterly Journal of Economics* 128 (4): 1499–545.

Duflo, E., R. Hanna, and S. P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

Duflo, E., M. Kremer, and J. Robinson. 2011b. "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya." *American Economic Review* 101 (6): 2350–90.

Dupas, P. 2011. "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 3 (1): 1–34.

Dupas, P. 2014. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." *Econometrica* 82 (1): 197–228.

Dupas, P., and J. Robinson. 2013a. "Saving Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 5 (1): 163–92.

Dupas, P., and J. Robinson. 2013b. "Why Don't the Poor Save More? Evidence from Health Savings, Experiments." *American Economic Review* 103 (4): 1138–71.

Eble, A., P. Boone, and D. Elbourne. 2017. "On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics." *World Bank Economic Review* 31 (3): 687–707.

Evans, D. K., and A. Popova. 2017. "Cash Transfers and Temptation Goods." *Economic Development and Cultural Change* 65 (2): 189–221.

Evans, D. K., and A. Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer* 31 (2): 242–70.

Feigenberg, B., E. Field, and R. Pande. 2013. "The Economic Returns to Social Interaction: Experimental Evidence from Microfinance." *Review of Economic Studies* 80 (4): 1459–83.

Field, E., R. Pande, J. Papp, and N. Rigol. 2013. "Does the Classic Microfinance Model Discourage Entrepreneurship among the Poor? Experimental Evidence from India." *American Economic Review* 103 (6): 2196–226.

Fujiwara, T., and L. Wantchekon. 2013. "Can Informed Public Deliberation Overcome Clientilism? Experimental Evidence from Benin." *American Economic Journal: Applied Economics* 5 (4): 241–55.

Gechter, M. 2016. "Generalizing the Results from Social Experiments: Theory and Evidence." Working Paper, Departement of Economics, Boston University. Available at: http://www.personal.psu.edu/mdg5396/Gechter_Generalizing_Social_Experiments.pdf.

Giné, X., J. Goldberg, and D. Yang. 2012. "Credit Market Consequences of Improved Personal Identification: Field Experimental Evidence from Malawi." *American Economic Review* 102 (6): 2923–54.

Giné, X., D. Karlan, and K. Zinman. 2010. "Put Your Money Where Your Butt Is: A Commitment Contract For Smoking Cessation." *American Economic Journal: Applied Economics* 2 (4): 213–35.

Glewwe, P., N. Ilias, and M. Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3): 205–27.

Glewwe, P., M. Kremer, and S. Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1 (1): 112–35.

Gneezy, U., K. L. Leonard, and J. A. List. 2009. "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society." *Econometrica* 77 (5):1637–64.

Hanna, R., S. Mullainathan, and J. Schwartzstein 2014. "Learning through Noticing: Theory and Evidence from a Field Experiment." *Quarterly Journal of Economics* 129 (3): 1311–53.

Hjort, J. 2014. "Ethnic Divisions and Production in Firms." *Quarterly Journal of Economics* 129 (4): 1899–946.

Jensen, R. 2010. "The (perceived) Returns for Education and the Demand for Schooling." *Quarterly Journal of Economics* 125 (2): 515–48.

Jensen, R. 2012. "Do Labor Market Opportunities Affect Young Women's Work and Family Decisions? Experimental Evidence from India." *Quarterly Journal of Economics* 127 (2): 753–92.

Jensen, R. T., and N. H. Miller. 2011. "Do Consumer Prices Subsidies really Improve Nutrition?" *Review of Economics and Statistics* 93 (4): 1205–23.

Karlan, D., R. Osei, I. Osei-Akoto, and C. Udry. 2014. "Agricultural Decisions after Relaxing Credit and Risk Constraints." *Quarterly Journal of Economics* 129 (2): 597–652.

Karlan, D., and M. Valdivia. 2011. "Teaching Entrepeneurship: Impact of Business Training on Microfinance Clients and Institutions." *Review of Economics and Statistics* 93 (2): 510–27.

Karlan, D., and J. Zinman. 2009. "Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment." *Econometrica* 77 (6): 1993–2008.

King, E. M., and J. R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *World Bank Research Observer* 24 (1): 55–84.

Kowalski, A. E. 2016. "How to Examine External Validity within an Experiment." *Mimeo.*, Available at: http://www.econ.yale.edu/~ak669/jep.latest.draft.

Kremer, M., J. Leino, E. Miguel, and A. Peterson Zwane. 2011. "Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions." *Quarterly Journal of Economics* 126 (1): 145–205.

Kremer, M., E. Miguel, and R. Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics* 91 (3): 437–56.

Levitt, S. D., and J. A. List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review* 53 (1): 1–18.

Lucas, A. M., and I. M. Mbiti. 2014. "Effects of School Quality on Student Achievement: Discontinuity Evidence from Kenya." *American Economic Journal: Applied Economics* 6 (3): 234–63.

Macours, K., N. Schady, and R. Vakis. 2012. "Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment." *American Economic Journal: Applied Economics* 4 (2): 247–73.

Macours, K., and R. Vakis. 2014. "Changing Households' Investment Behaviour through Social Interactions with Local Leaders: Evidence from a Randomised Transfer Programme." *Economic Journal* 124 (576): 607–33.

McCambridge, J., J. Witton, and D. R. Elbourne. 2014. "Systematic Review of the Hawthorne Effect: New Concepts Are Needed to Study Research Participation Effects." *Journal of Clinical Epidemiology* 67 (3): 267–77.

McKenzie, D., and C. Woodruff. 2013. "What Are We Learning from Business Training and Entrepeneurship Evaluations around the Developing World?" *World Bank Research Observer* 29 (1): 48–82.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, and R. Glennerster, *et al.* 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

Moffit, R. 2004. "The Role of Randomized Field Trials in Social Science Research." *American Behavioral Scientist* 47 (5): 506–40.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. 2010. "CONSORT 2010 Explanation and Elaboration: Updated Guidelines for reporting Parallel Group Randomised Trials." *BMJ* 340: c869.

Muller, S. M. 2015. "Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Experiments." *World Bank Economic Review* 29: S217–25.

Muralidharan, K., P. Niehaus, and S. Sukhtanker. 2017. "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India." NBER Working Paper No. 23838, National Bureau of Economic Research, Cambridge, MA.

Muralidharan, K., and V. Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *Quarterly Journal of Economics* 130 (3): 1011–66.

Muralidharan, K., and S. Venkatesh. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *Economic Journal* 120 (546): F187–203.

Muralidharan, K., and S. Venkatesh. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1): 39–77.

Olken, B. A., J. Onishi, and S. Wong. 2014. "Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia." *American Economic Journal: Applied Economics* 6 (4): 1–34.

Oster, E., and R. Thornton. 2011. "Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 3 (1): 91–100.

Pearl, J., and E. Bareinboim. 2014. "External Validity: From Do-Calculus to Transportability across Populations." *Statistical Science* 29 (4): 579–95.

Peters, J., J. Langbein, and G. Roberts. 2016. "Policy Evaluation, Randomized Controlled Trials, and External Validity–A Systematic Review." *Economics Letters* 147: 51–54.

Pradhan, M., D. Suryadarma, A. Beatty, M. Wong, A. Gaduh, A. Alisjahbana, and R. P. Artha. 2014. "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia." *American Economic Journal: Applied Economics* 6 (2): 105–26.

Prittchet, L., and J. Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review* 105 (5): 471–5.

Ravallion, M. 2012. "Fighting Poverty One Experiment at a Time: A Review of Abhijit Banerjee and Esther Duflo's Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty." *Journal of Economic Literature* 50 (1): 103–14.

Robinson, J. 2012. "Limited Insurance within the Household: Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 4 (4): 140–64.

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 W. Easterly and J. Cohen, 24–54. Washington, DC: Brookings Institution Press.

Roe, B. E., and D. R. Just. 2009. "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments, and Field Data." *American Journal of Agricultural Economics* 91 (5): 1266–71.

Roetman, E. 2011. "A Can of Worms? Implications of Rigorous Impact Evaluations for Development Agencies." 3ie Working Paper, Organisation International Impact Initiative (3ie), New Delhi.

Saretsky, G. 1972. "The OEO PC Experiment and the John Henry Effect." *The Phi Delta Kappan* 53 (9): 579–81.

Schulz, K. F., D. G. Altman, and D. Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMC Medicine* 8 (1): 18.

Simons, A. M., T. Beltramo, G. Blalock, and D. I. Levine. 2017. "Using Unobtrusive Sensors to Measure and Minimize Hawthorne Effects: Evidence from Cookstoves." *Journal of Environmental Economics and Management* 86: 68–80.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf. 2011. "The Use of Propensity Scores to assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (2): 369–86.

Tarozzi, A., A. Mahajan, B. Blackburn, D. Kopf, L. Krishnan, and J. Yoong. 2014. "Micro-loans, Insecticide-Treated Bednets, and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India." *American Economic Review* 104 (7): 1909–41.

Temple, J. R. W. 2010. "Aid and Conditionality." In *Handbook of Development Economics, vol. 5*, edited by P. Schultz and J. Strauss, 4417–511. Elsevier: North Holland.

Vicente, P. C. 2014. "Is Vote Buying Effective? Evidence from a Field Experiment in West Africa." *Economic Journal* 124 (574): F356–87.

Vivalt, E. 2017. "How Much Can We Generalize from Impact Evaluations?" Mimeo. Australian National University. Available at: http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf.

Voors, M., E. M. E. Nillesen, P. Verwimp, E. H. Bulte, R. Lensink, and D. P. van Soest. 2012. "Violent Conflict and Behavior: A Field Experiment in Burundi." *American Economic Review* 102 (2): 941–64.

Woolcock, M. 2013. "Using Case Studies to Explore the External Validity of 'Complex' Development Interventions." *Evaluation* 19 (3): 229–48.

Zwane, A. P., J. Zinman, E. Van Dusen, W. Pariente, C. Null, E. Miguel, and M. Kremer et al. 2011. "Being Surveyed can change later Behavior and related Parameter Estimates." *Proceedings of the National Academy of Sciences* 108 (5): 1821–6.

Appendix A: Reviewed Papers and Ratings

Author	**Question 1:** Participants aware of experiment/ study?	**Question 2:** Account for HJHE?	**Question 3:** Scaled-up program discussed?	**Question 4:** Long-run discussed?	**Additional Question:** Does the paper generalize?	**Question 5:** Policy population or restrictions discussed?	**Question 6:** Special care discussed?	**Question 7:** Implementa-tion partner?	**Author Response:** Feedback from the authors received?
Aker et al. (2012)	No	N/A	Yes	Yes	Yes	Yes	No	NGO	Yes
Alatas et al. (2012)	Yes	Participants are NOT aware	No	Yes	Yes	Yes	No	Government	No
Ashraf et al. (2010)	Yes	Yes	No	Yes	Yes	Yes	No	NGO	No
Ashraf et al. (2014)	Yes	No	No	Yes	Yes	Yes	No	Researcher	No
Attanasio et al. (2011)	No	N/A	Yes	No	Yes	Yes	No	Government	Yes
Baird et al. (2011)	Yes	Yes	No	No	Yes	Yes	No	NGO	No
Barrera-Osorio et al. (2011)	Yes	No	No	No	Yes	No	No	Regional Public Authority	Yes
Bertrand et al. (2010)	No	N/A	No	No	Yes	Yes	No	Firm	No
Björkman and Svennsson (2009)	No	N/A	Yes	Yes	No	N/A	Yes	NGO	Yes
Blattman et al. (2014)	Yes	No	Yes	Yes	Yes	Yes	No	Government	Yes
Blimpo (2014)	Yes	No	Yes	No	Yes	No	No	Researcher	Yes
Bloom et al. (2013)	Yes	Yes	No	Yes	Yes	Yes	No	Researcher	Yes

Author	**Question 1:** Participants aware of experiment/study?	**Question 2:** Account for HJHE?	**Question 3:** Scaled-up program discussed?	**Question 4:** Long-run discussed?	**Additional Question:** Does the paper generalize?	**Question 5:** Policy population or restrictions discussed?	**Question 6:** Special care discussed?	**Question 7:** Implementation partner?	**Author Response:** Feedback from the authors received?
Burde and Linden (2013)	No	N/A	No	No	Yes	No	No	NGO	No
Casey et al. (2012)	No	N/A	No	Yes	Yes	Yes	No	Government	No
Chinkhumba et al. (2014)	Yes	No	Yes	No	Yes	Yes	No	Researcher	Yes
Cohen and Dupas (2010)	No	N/A	Yes	Yes	Yes	Yes	No	Researcher	Yes
Collier and Vicente (2014)	No	N/A	No	No	Yes	Yes	No	NGO	Yes
Das et al. (2013)	No	N/A	Yes	Yes	Yes	Yes	Yes	NGO	Yes
de Mel et al. (2009a)	No	N/A	No	No	Yes	Yes	No	Researcher	No
de Mel et al. (2013)	Yes	No	No	No	Yes	Yes	No	Researcher	No
Drexler et al. (2014)	No	N/A	No	No	Yes	No	No	Firm	No
Duflo et al. (2011a)	No	N/A	Yes	No	Yes	Yes	Yes	NGO	Yes
Duflo et al. (2011b)	No	N/A	Yes	No	Yes	No	No	NGO	Yes
Duflo et al. (2012)	Yes	No	No	Yes	Yes	Yes	Yes	NGO	Yes
Duflo et al. (2013)	No	N/A	No	Yes	Yes	Yes	No	Regional Public Authority	Yes

Author	**Question 1:** Participants aware of experiment/study?	**Question 2:** Account for HJHE?	**Question 3:** Scaled-up program discussed?	**Question 4:** Long-run discussed?	**Additional Question:** Does the paper generalize?	**Question 5:** Policy population or restrictions discussed?	**Question 6:** Special care discussed?	**Question 7:** Implementation partner?	**Author Response:** Feedback from the authors received?
Dupas (2011)	No	N/A	Yes	Yes	Yes	Yes	Yes	NGO	Yes
Dupas and Robinson (2013a)	Yes	No	Yes	Yes	Yes	Yes	No	Firm	Yes
Dupas and Robinson (2013b)	Yes	No	Yes	Yes	Yes	Yes	No	Researcher	Yes
Dupas (2014)	Yes	No	Yes	Yes	Yes	No	Yes	Researcher	Yes
Feigenberg et al. (2013)	No	N/A	No	Yes	Yes	Yes	No	Firm	Yes
Field et al. (2013)	No	N/A	Yes	Yes	Yes	Yes	No	Firm	Yes
Fujiwara and Wantchekon (2013)	No	N/A	Yes	No	Yes	Yes	No	Researcher	Yes
Giné et al. (2010)	No	N/A	Yes	Yes	Yes	Yes	No	Firm	Yes
Giné et al. (2012)	No	N/A	No	Yes	Yes	Yes	No	Government	Yes
Glewwe et al. (2009)	No	N/A	No	No	Yes	Yes	No	NGO	Yes
Glewwe et al. (2010)	No	N/A	No	No	Yes	No	No	NGO	Yes
Hanna et al. (2014)	No	N/A	No	Yes	Yes	No	No	Researcher	No

Author	**Question 1:** Participants aware of experiment/study?	**Question 2:** Account for HJHE?	**Question 3:** Scaled-up program discussed?	**Question 4:** Long-run discussed?	**Additional Question:** Does the paper generalize?	**Question 5:** Policy population or restrictions discussed?	**Question 6:** Special care discussed?	**Question 7:** Implementation partner?	**Author Response:** Feedback from the authors received?
Jensen (2010)	No	N/A	No	Yes	Yes	Yes	No	Researcher	Yes
Jensen (2012)	No	N/A	No	No	Yes	Yes	No	Researcher	Yes
Jensen and Miller (2011)	No	N/A	Yes	No	Yes	No	No	Government	Yes
Karlan et al. (2014)	Yes	Participants are NOT aware	Yes	No	Yes	Yes	Yes	Government	No
Karlan and Valdivia (2011)	No	N/A	Yes	No	No	N/A	Yes	NGO	No
Kremer et al. (2011)	No	N/A	No	No	Yes	No	No	NGO	No
Kremer et al. (2009)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	NGO	Yes
Macours et al. (2012)	No	N/A	No	No	Yes	Yes	No	Government	No
Macours and Vakis (2014)	Yes	No	No	No	Yes	No	No	Government	No
Muralidharan and Venkatesh (2010)	Yes	Yes	No	No	Yes	Yes	No	NGO	No
Muralidharan and Venkatesh (2011)	No	N/A	Yes	Yes	Yes	Yes	Yes	NGO	No
Olken et al. (2014)	No	N/A	Yes	Yes	Yes	Yes	No	Government	Yes

Author	**Question 1:** Participants aware of experiment/study?	**Question 2:** Account for HJHE?	**Question 3:** Scaled-up program discussed?	**Question 4:** Long-run discussed?	**Additional Question:** Does the paper generalize?	**Question 5:** Policy population or restrictions discussed?	**Question 6:** Special care discussed?	**Question 7:** Implementation partner?	**Author Response:** Feedback from the authors received?
Oster and Thornton (2011)	No	N/A	No	No	Yes	Yes	No	Researcher	Yes
Pradhan et al. (2014)	No	N/A	No	No	Yes	No	No	Firm	Yes
Robinson (2012)	Yes	No	No	No	Yes	Yes	No	Researcher	Yes
Tarozzi et al. (2014)	No	N/A	Yes	No	Yes	Yes	Yes	Firm	Yes
Vicente (2014)	No	N/A	No	No	Yes	Yes	No	Government	Yes

Appendix B: Excluded Papers and Reason for Exclusion

Author	Reason for exclusion
Adhvaryu (2014)	Quasi-experiment
Armantier and Boly (2013)	Artefactual experiment
Ashraf (2009)	Behavioral Field experiment
Attanasio et al. (2012)	Artefactual experiment
Bauer et al. (2012)	Behavioral Field experiment
Beaman and Magruder (2012)	Artefactual experiment
Beaman et al. (2009)	Natural experiment
Besley et al. (2012)	Theoretical paper
Bursztyn and Coffman (2012)	Natural experiment
Cai et al. (2009)	Behavioral field experiment
Chassang et al. (2012)	Theoretical paper about RCTs
De Mel et al. (2009b)	Reply to a previously published article
DiTella and Schargrodsky (2013)	Natural experiment
Gneezy et al. (2009)	Artefactual experiment
Hjort (2014)	Natural experiment
Karlan and Zinman (2009)	Behavioral field experiment
Lucas and Mbiti (2014)	Quasi-experiment
Voors et al. (2012)	Artefactual experiment