Practical Challenges and Strategies for Randomized Control Trials in Agricultural Extension and Other Development Programs

Michael J. Culbertson
Daniel McCole
Paul E. McNamara

This manuscript has been published in the *Journal of Development Effectiveness* (2014), v 6, no 3, pp 284-299, doi:10.1080/19439342.2014.919339

Abstract

Randomized Control Trials (RCTs) can yield information about the effectiveness of agricultural extension and other development programs, facilitating cost-benefit analyses and policy decisions under resource constraints. However, even after RCT design questions have been settled, a number of other practical challenges to successful RCTs remain. As a guide to those interested in applying RCTs more extensively, we outline several ethical, organizational, design, and field-based challenges for RCTs, along with potential strategies for mitigating the challenges. We provide illustrations from our experience of an RCT of the Community Knowledge Worker program, a novel agricultural extension model in Uganda.

Keywords

Randomized control trial (RCT), Impact evaluation, Agricultural extension, Uganda, Africa.

Acknowledgement

This paper was made possible by the generous support of the American people through the United States Agency for International Development (USAID). The authors' views expressed in this publication do not necessarily reflect the views of the USAID or the US Government.

I. Introduction

Recent years have seen increasing frustration with development assistance, as many on both the giving and receiving ends of international aid perceive that decades of effort have produced considerably smaller results on the whole than originally envisioned. The question of what, exactly, development projects have accomplished has led many to call for more extensive evaluation, particularly impact evaluation, in international development. While impact evaluation is becoming more common in social sectors, it is relatively rare among agricultural extension programs—though not because the effects of extension programs are better known than effects of programs in other sectors (Winters, Maffioli, & Salazar, 2011). The push for more impact evaluation has not been received uncritically, and much of the debate about impact evaluation centers on a key study design element: whether to randomize assignment to the program of interest. The contention over Randomized Control Trials (RCTs) may be stimulated by prominent governmental and funding agencies that have emphasized the utility of RCTs to such a degree as to suggest (whether actually or seemingly) that all other evaluation designs are inferior and less trustworthy. A more reasonable approach to evaluation design is to let the appropriateness of particular methods given evaluation goals, information needs, and the state of program development guide design decisions, including the decision to randomize (e.g., Patton, 2008). If randomization yields a unique insight apart from other methods, agricultural extension programs and other development efforts could benefit from expanded judicious use of RCTs. What, then, does randomization provide for an evaluation?

(a) Case for RCTs

Governments and development funding agencies look for evaluations of programs and policies in order to determine whether funded activities are meeting the organization's stated aims—in other words, to determine if decisions have the desired effects. But, decision-makers (such as program staff, funding officers, legislators, and administrative policy-makers) often need to know more than whether a particular option will produce a positive outcome: After all, need usually outpaces resources, and in order to promote good stewardship, decision-makers try to spend their limited resources on the programs that will achieve the most good. Making these decisions requires knowledge not only of *whether* a program achieves an end, but also of the *extent or size* of the program's effects. While there are a number of ways to examine the relationship between causes and effects, RCTs are often particularly well suited for answering the specific question "How much of an observed effect is attributable to this particular cause (the program or policy)?" Randomized assignment is so amenable for this question because it eliminates the validity threat of selection bias, which occurs when those who choose to participate in a program are already more likely to have better outcomes for reasons entirely other than their participation in the program (Henry, 2009).

Randomization is not a "silver bullet," and though RCTs eliminate selection bias, they can still fall prey to a host of other validity threats, just as any other research study (Bickman & Reich, 2009; Conrad & Conrad, 1994). Moreover, RCTs typically only answer one evaluation question of interest—effect size—leaving other important evaluation questions unexplored. Other important evaluation questions typically unexplored by RCTs include examining the internal efficiency of program activities, understanding how programs achieve their effects, and determining whether program activities meet beneficiaries' most pressing needs or goals, among others. However, the fact that RCTs do not answer all evaluation questions is no reason to pass over randomization when it is methodologically appropriate and the question of effect size is important for decision-making. In fact, mixed-methods evaluations that combine a randomized experiment with case study or theory-based process studies can provide much stronger evidence about a program than evaluations based on a single approach (Cook, 2000).

Agricultural extension is tightly linked with both government activity and development efforts targeting the poor around the world. In an age of tight budgets, decision-makers in these agencies need information about which techniques will make most economical use of limited resources. Refining extension activities will require more evaluations of all types, including RCTs where appropriate.

As in any research, deciding that conditions are right for an RCT is just the first step in a process of design and implementation. Others have discussed extensively how to set up randomization schemes and build support among stakeholders (Bickman, 1985; Boruch & Wothke, 1985; Dennis & Boruch, 1994; Duflo, Glennerster, & Kremer, 2007); but even when the design is settled, practical challenges to the successful completion of an RCT remain. RCTs of agricultural extension programs can be particularly challenging due to their developmental context, but the successful completion of extension RCTs in the past suggests that these challenges are not insurmountable (e.g., Buck & Alwang, 2011; Duflo, Kremer, & Robinson, 2010; Stringer et al., 2011). This paper outlines a number of ethical, organizational, design, and field challenges that can arise during the implementation of RCTs in developing nations, along with potential strategies for mitigating the challenges and illustrations from an ongoing RCT of an agricultural extension program in Uganda.

(b) Context

The Grameen Foundation's Community Knowledge Worker program aims to increase the availability of agricultural information for poor rural farmers in Uganda. The program equips model farmer community members with a smartphone and trains them how to use a pre-loaded smartphone application to access a database of continuously updated agricultural information. So trained, these Community Knowledge Workers (CKWs) serve as liaisons between their neighbors and new insight into farming. The database includes information about farming tips and best practices, agricultural service

providers, weather forecasts, and markets, including current prices. Unlike many other innovative agricultural extension initiatives based on information and communication technologies (ICT) such as radio or SMS text messaging, CKWs are not simply passive conduits of information; rather, they actively interpret and teach their neighbors based on the information they obtain by searching the agricultural database, and the best CKWs actively seek out neighbors who could benefit from the program and follow-up on the provided information at a later date. As such, CKWs serve as community-level agricultural extension workers, with less formal education than traditional extension officers, but with a much improved officer-to-farmer ratio. Moreover, the CKW technology platform also includes an electronic survey enumeration program, and CKWs are trained and serve as enumerators to collect information about farmers for Grameen and other development organizations interested in providing services in the same area. Thus, CKWs provide a two-way flow of information between local farmers and service providers.

When we began working with the Grameen Foundation (GF), the CKW program had been operational for two years and had expanded to 17 districts across Uganda, with about 800 CKWs serving over 70,000 farmers. GF always rolls out the CKW program in partnership with another development organization; in addition to the general CKW training, CKWs in any given area are trained to service the partner's development program as well, either through surveys or a given emphasis in development messages, for example, concerning partner services or agricultural information. The next expansion of the CKW program was planned with the East Africa Dairy Development (EADD) in the vicinity of Masaka. EADD establishes and supports the development of dairy hubs consisting of local dairy farmer associations that support farmers through training, milk bulking, and providing other dairy business services. CKWs would operate in areas served by EADD dairy hubs, providing their usual agricultural information services and encouraging farmer participation in dairy hub activities. The expansion of the CKW program into the Masaka area would be gradual, with an initial deployment to a small number of hubs immediately, using leftover funds from EADD's last grant cycle, and a second, larger deployment to the remaining hubs once the new grant cycle began.

Based on staff experience and a retrospective difference-in-differences evaluation (Van Campenhout, 2012), GF believed that the CKW program was providing positive impact for farmers. GF was interested in expanding the CKW program, both in Uganda and internationally. In order to convince policy makers of the benefits of adopting the CKW model, GF wanted to obtain strong causal estimates of the size of the CKW program's impact, which could then be incorporated into a cost-benefit argument. This goal and the planned expansion of the program into a new district provided conditions amenable to a randomized control design.

2. Ethical Challenges and Strategies

Some critics of RCT-based evaluation point to the ethical implications of randomizing access to potentially beneficial programs. While it is certainly very important to consider the ethics of an RCT design for each particular evaluation scenario, the ethical challenges are often not insurmountable. Ethical challenges typically concern the justification for denying services to those in need (Conner, 1980). If there are resources to support all eligible participants, how can one withhold beneficial services from some, while extending them to others? But, when resources are limited, denying services to some eligible participants becomes a necessity, and the question shifts to how to choose whom to deny. In one sense, lotteries for scarce resources (as in RCTs) are fair, since all eligible participants have the same chance of gaining access. However, lotteries are also completely blind to differences between eligible participants, treating the most needy the same as the less needy.

Randomized assignment is not suitable in all circumstances, but some conditions decrease the ethical concerns with denying services randomly. For example, if limited resources are not allocated based on need prioritization, there is little ethical difference between randomized allocation and allocation on a "first come" basis (in fact, a randomized allocation could be considered more ethical, as it increases access for participants who would not be able to reach services quickly enough). Similarly, in many cases, it is not feasible to implement a new program over a large area all at once, and a gradual rollout is a practical necessity. When a time-delayed rollout is already planned, choosing the order of the rollout randomly to permit an RCT design does not delay access to resources beyond the original plan. Denying access to a new program may also be permissible when eligible participants already have access to other similar services. Moreover, evaluation designs that test a new service against existing services (instead of a service against no service) not only address the denial of service issue, but also provide potentially more-useful information to policy makers.

Even when conditions are amenable to randomizing access to services, it may be preferable to include additional value for the control group. For example, program staff could commit to providing the control group with additional services beyond the standard intervention (e.g., a "catch-up" investment) at the end of the study in order to compensate for lost time relative to the treatment group. Moreover, if frequent monitoring of effects is feasible, it may be possible to end the RCT early, as soon as sufficient evidence of the size of the positive impact has been detected. However, this strategy should be used with caution (Adhikari et al., 2005; Bassler, Montori, Briel, Glasziou, & Guyatt, 2008): Even when there is no true effect, each statistical test runs the risk of detecting an "effect" by chance (Type I error). If a series of tests is long enough, some test will eventually show an effect just by chance, potentially yielding faulty conclusions about program impact. Finally, for interventions that promote structural or organizational development, an extra investment

fund can be established equivalent to what the control group would have received over the course of the study, delivered at the end of the study to help "catch up" the control group's development.

GF and EADD did not have the resources to implement the CKW program in all of Masaka at once, and the rollout in the Masaka area was already planned to occur in phases. The intention was to provide CKWs to all of the Masaka-cluster dairy hubs eventually, but the order would be determined randomly in order to permit an impact evaluation with RCT design. In fact, we were able to argue for a greater initial investment in the Masaka area than was originally planned on grounds of the evaluation design, bringing services to some farmers in the area sooner. We also agreed to a conservative early-termination rule to minimize the delay for hubs in the control condition.

3. Organizational Challenges and Strategies

(a) Limited technical capacity

Since development organizations generally focus on designing and implementing their development programs, they may have limited technical capacity for designing and conducting impact evaluation, particularly in smaller organizations. Even when an organization collaborates with trained external evaluators, this limited capacity can present several challenges during evaluation design. Not fully appreciating the merits of different evaluation methods, organizational administrators may insist on an RCT design because they believe it is the impact evaluation "gold standard," even if other designs are more appropriate for the program. Moreover, program staff may not realize that RCTs require random assignment to treatment, thinking that random sampling for a survey would be sufficient. Finally, program staff may not have clearly articulated the desired or expected program impact, how the program achieves its impact, or how long it may take for measurable impact to become apparent. Without this clarity, program staff may have unrealistic expectations about how an RCT for their program would take place, what it would measure, and how long it would last, potentially leading to friction during the design process.

Strategies. Early, clear communication is key for bridging capacity gaps. When an organization asks for an RCT, evaluators should ensure that staff understand what an RCT entails, as well as the conditions in which RCTs are appropriate and likely to be beneficial. Evaluators may have to provide education on basic evaluation principles and the merits of different evaluation designs. If the program goals, theory of change, expected impact, and evaluation objectives are not clearly defined, evaluators should first lead program staff through self-reflective exercises to describe the state of the program before thinking about RCT design questions, including whether an RCT is most appropriate. However, it is also important for evaluators to resist the temptation to bypass key stakeholders during design decision-making in an attempt to complete the design quickly—stakeholders have a deep

familiarity with the program and context that is crucial for arriving at a sound evaluation plan. Fluid, two-way communication blends technical expertise and local knowledge about the program to assess the feasibility of different evaluation design options.

(b) Divergent objectives

When development organizations collaborate, program partners may have different programmatic goals, though they share common means. Consequently, a meaningful treatment definition for one partner may not be meaningful for another. Moreover, partners may place different value on evaluation evidence according to their divergent objectives. Since RCTs place constraints on program implementation, all partners need to back the evaluation plans and believe that the study will provide benefits for their own organization to outweigh the opportunity cost of participation.

EADD's main programmatic goal was the development of self-sufficient dairy hubs that supported dairy farming in the surrounding region through training, access to resources, and milk bulking and marketing. Based on past partnership with GF in other areas in Uganda, EADD believed that the CKW program was instrumental in driving the farmer participation in their dairy hubs necessary for the hubs to develop to self-sufficiency. Since EADD's goal was the development of the dairy hubs, not the development of the CKW program per se, EADD was initially reluctant to participate in an RCT comparing their program without CKWs to their program with CKWs. The information provided by such a contrast simply was not valuable enough for EADD to give up the choice of where to place CKWs, which EADD felt could potentially jeopardize their ability to meet their program commitments.

Strategies. Even if an RCT is intended primarily to evaluate only one organization's role in a program, having all programmatic partners at the table early in evaluation discussions is crucial for working toward a mutually beneficial evaluation design. With all partners in the conversation, the evaluation team can develop a shared understanding of all partners' goals and evaluation needs, encourage ownership of the evaluation design, and identify treatments meaningful for all partners. This may require having more than two treatments (program/control) to distinguish features relevant for different partners.

EADD staff knew well that the success of their dairy hubs required more than only access to information—for example, if CKWs teach farmers that applying flea dip to cattle reduces disease, but farmers cannot access or afford the dip, the long-term viability of the dairy hub is threatened, since the hub's success depends on farmers' dairy production. EADD suggested adding a treatment to test the effect of helping dairy hubs to establish an agro-vet shop that would give farmers greater access dairy-related agricultural inputs. The resulting three-treatment design (dairy hubs alone, dairy hubs with CKWs, dairy hubs with CKWs and agro-vet shop) provided contrasts that informed the programmatic goals of both GF

and EADD and was instrumental in the negotiation of a satisfactory compromise for the evaluation plan.

(c) Program commitments

Following Donald Campbell's admonition to evaluate only "proud programs" (Campbell, 1984), an RCT should come after an innovative program is well established. However, establishing development programs often requires making a variety of commitments. Entrance into a community may require commitments to provide certain levels of service or to keep a particular implementation schedule. Access to grant funds may require commitments to funders to deliver rapid or expansive results. Random assignment in an RCT can threaten these programmatic commitments.

As a condition of funding, EADD had committed to achieving self-sufficiency of a certain number of hubs during their funding cycle. Since they viewed CKWs as critical to hub development, they were reluctant to take the chance that the hubs closest to self-sufficiency would be randomly assigned not to receive the CKWs that would provide the final boost of development to achieve commitments to their funder.

Strategies. Program commitments are less likely to inhibit RCT design when program planning considers evaluation from the beginning. If program planners have in mind that an RCT may be desirable in the future, they may be able to craft programmatic commitments loosely enough to accommodate randomization. In other cases, it may be possible to negotiate flexibility with commitments to funders if the possibility for an RCT arises unexpectedly. Funding agencies are increasingly looking for strong evidence of program impact, and they may be willing to allow flexibility in output monitoring in exchange for the evidence of outcomes and impact provided by the RCT. Finally, when commitments are not malleable, it may be possible to restrict the RCT study region to accommodate the constraints.

In order to participate in the RCT and still make their funding commitments, EADD proposed that two of the twelve initial dairy hubs be excluded from the study. EADD felt that these two hubs had the best chance of becoming self-sufficient by the end of the funding cycle. The two hubs set aside would receive CKWs as originally planned, but would not be considered in the impact evaluation; and the remaining ten hubs, plus two nearby replacement hubs, would be randomly assigned according to the study design.

(d) Decentralized management

Particularly in organizations that work over large rural areas, staff in the organization's main office may be less familiar with program activity than workers on the ground. Programs often morph between planning and implementation stages as imagined context meets reality and staff discover that plans need to be adapted to account for local conditions or unforeseen circumstances. These decisions sometimes happen quickly by site

managers, and main-office staff may simply not be aware of all the decisions remote managers are making.

Staff in EADD's Kampala office knew the location of their dairy hub offices and the manager or contact person for each hub, and they knew generally the kinds of hub activities; however, the central office did not keep track of all hub activities, and consequently could not specify the exact region of activity around the hub office. This may have been in part because EADD planned for hubs to develop toward autonomy, with EADD taking only a supporting role. Since the hub contacts were not easily reached by phone, this rendered planning for baseline data collection difficult, as in several cases we had to wait until enumerators were on location to meet with hub personnel before we could perform the first (village-level) stage of survey sampling.

Strategies. For decentralized programs, include site-level managers early in the evaluation planning process. Not only will they be able to provide key logistical details that may be unavailable by higher-level staff, they can alert the evaluation team to differences between the program as planned and implemented. If possible, it may also be helpful to observe the program in action at several different sites to establish a sense of site-specific adaptations or environment that may affect the meaningfulness of evaluation questions, methods, and instruments.

(e) Internal resources and external credibility

When an organization has high research capacity, it may be tempting to leverage internal resources for data collection or analysis in order to reduce costs. However, in some cases, using internal resources can cause a conflict of interest, threatening the external credibility of the evaluation results. For example, CKWs are trained survey enumerators; using CKWs from an earlier implementation region to collect data for the RCT could have reduced data collection costs. However, we were concerned about the potential for a subconscious bias introduced by direct beneficiaries of a program collecting data on the program's effectiveness, which could render the results suspect to others.

Strategies. Generally, leveraging internal resources can render an RCT more feasible; however, it is important to think carefully about the potential bias that could be introduced. Make use of internal resources where the threat of bias is low and independence is not necessary for the credibility of evaluation results. We chose to make use of GF's existing CKW technology platform for mobile electronic data collection, since the computer software and servers were neutral with respect to evaluation results. However, we opted for specially trained external enumerators instead of CKWs to limit any hopeful bias toward positive evaluation results.

4. Design Challenges and Strategies

(a) Group equivalence and statistical power

In order to isolate the effect of the treatment from the effects of other contextual variables, RCTs use random assignment to ensure that the treatment and control groups are equivalent on all variables other than treatment assignment, *on average*. Moreover, the typical statistical tests covered in introductory statistics courses, such as the two-sample *t*-test for comparing two population means, are based on the variability of the *sample mean*. However, in community-, site-, and group-based interventions, only a very limited number of units (communities, sites, or groups) are available for random assignment. Since large samples are generally necessary to ensure that variability in averages is small, it can be difficult to ensure that treatment assignment is not correlated with some contextual variable, by chance, and that the standard error of sample means will be small enough to separate the treatment effect from variability in the population.

Strategies. Collecting baseline data can increase statistical power and reduce the threat of accidentally correlated contextual variables. Variability in the rate of change of outcome measures (e.g., increase in income over two years) is generally less than variability in the absolute outcome levels (e.g., income levels in a diverse community). In addition to increasing power, controlling for baseline outcome measures also controls for any other contextual variable that may be correlated with the outcome measure at the beginning of the study, leaving only a smaller set of rate-related variables that may be accidentally correlated with both treatment and outcomes. Moreover, when all of the study units are known ahead of time (i.e., enrollment in the study is not ongoing), statistical power can be increased further by optimally matching unit pairs on relevant outcome variables and covariates and randomizing within each pair (Greevy, Lu, Silber, & Rosenbaum, 2004; Raudenbush, Martinez, & Spybrook, 2007), similar to blocking in classical experimental design.

Finally, the sampling process is not the only source of variability for hypothesis testing in RCTs: The process of random assignment can also be used to construct statistical tests (e.g., Rosenbaum, 2002; Raab & Butcher, 2005; Rader, 2011), called randomization, permutation, or exact tests. Typically, randomization inference tests Fisher's "sharp" null hypothesis that the treatment has no effect for *any* of the *given* units (subjects), unlike sampling-based tests, which hypothesize no *average* treatment effect in the *population*. As such, the inference is only to the study sample, not to the population at large. This has the advantage of removing sampling variability from the test, but the disadvantage of limiting the scope of inference. However, although we are usually interested in effects in the entire population, not in a particular sample, knowing that an intervention has an effect for the particular individuals (communities, groups, etc.) under study can be the first step in constructing an argument for effects in larger populations (Rosenbaum, 2010, sec. 2.6).

In addition to incorporating baseline data collection and randomization tests into the design of the CKW program evaluation, we made use of matching before treatment assignment to control hub variability. We identified 21 variables of interest from the baseline survey, including the most important outcome variables, income, telecommunications, land use, and key agricultural indicators. We also included the geographic size and population density for each hub from census data. Using hub-level averages of the household data, we calculated the Euclidean distance (sum of squared deviations after standardization) between each pair of hubs and grouped the twelve hubs into four triads that minimized the intra-group differences. Within each group, one hub was randomly assigned to each of the three treatments.

(b) Spillover and maintaining assignment

Often for RCTs that randomize over individuals, maintaining strict random assignment can be challenging. If program staff perform assignment, some may be tempted to make a special exception for a particularly needy case. Individuals dissatisfied with their treatment assignment may maneuver themselves into another treatment group, or may drop out of the study altogether. In site-randomized trials, individuals may travel to another site to participate in the desired treatment, if they believe the program at that site may be more beneficial for them. Moreover, if elements of the treatment spread easily, as with knowledge-based interventions , which are common in agricultural extension programs, individuals in the control group may benefit from nearby treatment areas by word-of-mouth. For example, if a CKW teaches a farmer in the treatment group how to stop the spread of banana wilt, the farmer may call fellow farmers known to experience banana wilt, who may be in the control group, to inform them of what was learned. In general, one might want the benefits of program activity to spillover to non-participants in order to extend the program's impact, but these spillover effects are at cross-purposes with identifying the size of the program's impact when they spread between treatment groups.

Strategies. When possible, restricting the random assignment procedures to a limited number of staff can reduce the chance of exceptions. In all cases, developing organizational buy-in for the study and method is key—staff need to fully understand the purpose and importance of random assignment for the validity of the study results (Stevens, 1994). Community-level interventions, though they usually have lower statistical power, can reduce the threat of ignored assignment, particularly if the study communities are not too close to one another, since the program either is or is not implemented in the community (Newman, Rawlings, & Gertler, 1994). When there is some flexibility in the size of the randomization units (e.g., individual, village, county, region), choosing a larger unit may reduce the threat of spillover, since individuals would have to travel farther to participate in the program and individuals in the treatment and control regions may be less likely to know one another the farther apart they are (and thus less likely to share new

knowledge). If treatment regions are contiguous, it may be helpful to place a buffer region around the boundaries in which no data are collected to reduce the measurement of individuals who received spillover effects. Finally, when spillover is likely, as with knowledge-based interventions such as in many agricultural extension programs, measuring proxies for spillover can provide statistical controls to reduce bias, and embedding the experiment in a rigorous quasi-experimental design provides a backup if randomization fails catastrophically (e.g., Dennis, 1990).

CKWs typically operate over a single Ugandan parish (around 10 square km). We knew from discussions with CKWs that farmers from neighboring parishes sometimes come to request information services. Before the EADD hub structure was known, we initially decided to randomize not at the parish level (individual CKWs), but at the sub-county level (a collection of 4-6 parishes) to minimize this spillover. (EADD hubs, which served as the randomization unit in the final design, are also about the size of a sub-county.) Moreover, the CKW administrative logs contain latitude and longitude coordinates for CKWs' homes and all interactions, including the evaluation survey responses—computing the distance from respondents in the control region to the nearest CKW provides one proxy for spillover effects due to traveling to receive services from CKWs in an adjacent sub-county or knowing an individual in the treatment area who could share new knowledge. The survey questionnaire also measures which respondents have had contact with a CKW in both control and treatment areas and includes a rough measure for how much respondents share farming information with others, both locally and remotely. In a survey of early participants in the program, about 60% of respondents who received services from a CKW had shared their new knowledge with someone else, but less than 3% of those who shared reported that they had shared with someone from another sub-county, suggesting that contamination effects due to word-of-mouth spillover across treatment regions are minimal in this study.

(c) Program fidelity and the black box

Multi-site programs may have considerable variation in program implementation. If all program sites are treated as equivalent, variation in the intensity of program implementation will likely reduce the average treatment effect obtained in an RCT. By itself, an RCT measures only whether a treatment has an effect and the size of that effect, not how or why the treatment produces its effect—the treatment is taken as a "black box." But, not knowing how or why a program works makes transferring the program to new areas more challenging, as it is more difficult to determine whether a key component of the program's success depends on some local contextual factor. For both of these reasons—measuring treatment intensity and understanding the "how" behind casual effect estimates—RCTs should incorporate concurrent studies of program implementation. Program evaluations should generally only undertake an RCT when there is good reason to believe that the

program has identifiable effects, and this may mean that the program process is already understood in part. However, even when there is good reason to anticipate a program's effects, an RCT can sometimes fail to detect them due to unforeseen complications. Incorporating a concurrent study of the program's inner workings often involves only marginally greater costs, affords opportunities to discover unanticipated positive or negative effects of the program, and provides useful information to program staff and others interested in implementing similar programs elsewhere while hedging the research investment of an RCT against failure by yielding insights even when no causal effects are detected.

Strategies. In addition to identifying expected program outcomes and impact, the evaluation team and program staff should think carefully through the program's theory of action—how, exactly, do the program inputs and activities achieve the outcomes? Sometimes program developers document such a theory of action during program design. Identify key points in the theory and plan to observe or measure whether the program is functioning as expected. If resources permit, incorporating qualitative data on program implementation (what happens, how participants respond to the program, what effects do participants attribute to the program and why) provides a rich base from which to craft descriptions of the mechanism behind RCT causal effect results. Finally, include measures of program fidelity (Century, Rudnick, & Freeman, 2010; Mowbray, Holter, Teague, & Bybee, 2003), implementation, or participation to account for variation in the treatment.

CKW administrative data provide information about the intensity of program implementation in different regions: The CKW Search application automatically logs every database search CKWs conduct, and CKWs register farmers before providing their first information service. The questionnaire for the impact survey includes measures of both intermediate (knowledge gain, new farming activities) as well as long-term (income, food security, poverty measures) impact, according to the program theory. Moreover, in addition to the main impact survey, a smaller concurrent survey of 200 farmers who have interacted with a CKW will probe more deeply the nature of the CKW-farmer interaction and how farmers respond to the information they receive.

(d) Eagerness to innovate

People who design development programs can be energetic and innovative and may have a stream of new ideas for improving the lives of others. While it may be tempting to implement these new ideas as soon as the details are worked out, changing the program design significantly during the course of an RCT makes interpreting the effects challenging—is the impact due to the program before the change or after the change? The challenge to maintain programmatic discipline increases when interventions involve multiple partner organizations, each of which may be generating new ideas to support the well being of program participants.

Strategies. Early and frequent communication about the importance of a stable program definition for the duration of the study is crucial, particularly if organizations are only beginning to develop evaluation capacity. Evaluators should probe program developers and staff about upcoming innovations and new ideas in development, and they should encourage innovations in regions outside the RCT study area, where possible. Particularly in partnerships, monitoring of program activity can identify and potentially correct for any accidental deviations in the study design. When any changes that do occur are carefully documented, it may be possible to incorporate programmatic changes into the analysis, particularly for programs with continuous enrollment (e.g., Dennis, 1990).

When we began the CKW evaluation, GF was in the late stages of planning a new micro-credit program to be deployed via CKWs. GF expected to implement the program within a year, but it would not be ready for deployment before training of the RCT CKWs began. We cautioned that implementing the new micro-credit program in the study area would weaken interpretation of the RCT results, as it would not be possible to separate the effects of the knowledge component of the CKW program from the new micro-credit program.

(e) Encroaching interventions

Development activities often do not occur in isolation—different organizations with different missions and activities often work in the same region, with varying levels of coordination. As long as other organizations' work is implemented evenly throughout the study area, or at least is not correlated accidentally with the RCT treatment assignment, this activity will not bias RCT results. However, particularly in place-based RCTs with a small number of randomized units, the effects of the program of interest may be obscured if another organization rolls out a new program in the RCT control area.

Strategies. While it is not generally possible to prevent other organizations from launching new programs that conflict with an ongoing impact evaluation, it may be possible to control for the effect of other programs in analysis. During RCT planning, determine which other development organizations are present and active—this information can also be incorporated into pre-randomization matching, if applicable. Then, as the RCT unfolds, monitor and record any major interventions that take place in the study area, such as new educational programs or asset transfers (e.g., animals, equipment, cash). Occurrence of major interventions can be controlled statistically through covariate adjustment. Geographic diversity in the study can help minimize risk of obscured effects due to external interventions, since it is less likely that other organizations are operating over exactly the same geographic area.

5. Field Challenges and Strategies

(a) Cultural and linguistic knowledge

While cultural competence is important for all evaluators (American Evaluation Association, 2011; SenGupta, Hopson, & Thompson-Robinson, 2004), an outsider can be only but so aware of cultural and linguistic issues. Since culture affects all aspects of research, including survey instruments, the meaningfulness of impact measures, and data collection procedures, deep knowledge of the culture is crucial for valid inferences. It may also be the case that program participants come from more than one culture, and cultural differences could affect program effectiveness.

Strategies. Ideally, the evaluation team would represent the cultures of program beneficiaries. When evaluators come from another culture, including a cultural insider on the team can be an invaluable resource, as the insider may notice issues that would never occur to even the most culturally sensitive outsider. The program theory, impact measures, survey instruments, and data collection procedures should be vetted with cultural insiders, who may be on the evaluation team, program staff, or program participants. Local enumerators can also alert evaluators to potential issues with data collection procedures and survey instruments.

Although most of the people in the CKW RCT study area were Baganda, the majority ethnic group in Uganda, the sub-county Sembabule included other ethnicities. When recruiting survey enumerators, the GF internal evaluator, who was a native Ugandan, attended to enumerators' accents. Even though most interviews were conducted in the majority language Luganda, several enumerators familiar with Luganda dialects influenced by the local language in Sembabule were recruited to facilitate communication with people of these ethnicities.

(b) Developing world infrastructure

Infrastructure limitations can pose logistical challenges that delay data collection or increase costs. For example, power and telecommunications systems may not be consistently available, inhibiting computer-based activities and coordination of teams. In rural settings in particular, transportation is by far the greatest limiting factor, which can constrain study design. Villages or households may be quite dispersed and roads in between of low quality, limiting the geographic range over which it is feasible to collect data and increasing time, vehicle, and fuel costs.

Initially, we planned to randomly assign the gender matches between interviewers and interviewee, since there has been some evidence of gender effects in survey responses (Benstead, 2010; Huddy, Billig, Bracciodieta, Moynihan, & Pugliani, 1997; Miller, Zulu, & Watkins, 2001). We composed enumerator teams with half men and half women, and each household would be randomly assigned either the husband or wife. If a particular household had only an unmarried or widowed man or woman, enumerators would swap with another

team member if the gender matching was not as assigned. However, we quickly discovered that we could only visit each village once, leaving no time for enumerators to swap households or to return to a household with no one at home. Due to the timing and transportation constraints, we had to drop strict random assignment of gender matching. Instead, we instructed enumerators to balance the number of male and female interviewees in each village as best as possible.

Strategies. When planning data collection logistics, build in slack and backup procedures for power or communication outages. This may include electrical generators for crucial functions and logistical plans that rely on minimal communication in the field. If time and budget allow, sending an advance team to map out rural terrain, locate selected villages, plan routes, and note road quality can be immensely helpful for developing an efficient logistical plan. Finally, vehicles suitable for low-quality roads are generally more expensive, but they can reduce delays due to vehicle break down. Ensure that budgets plan for vehicles well matched to the anticipated terrain.

(c) Trust

Unless an RCT is testing an addition or modification to an existing program, RCTs usually involve data collection in areas in which an organization does not already have a presence—areas in which the organization will soon be present and areas that will serve as comparison groups. Since the organization may be unknown, local stakeholders may not initially trust survey enumerators. Stakeholders may also be less interested in participating if they have experienced a "revolving door" of development organizations and researchers or if the region suffers from political tensions.

For example, the sub-county Sembabule had been experiencing ethnic tensions, including encroachment on historical tribal lands. Since there had additionally been recent land and cattle grabbing, some residents quickly became suspicious when outsiders (Ugandans from Kampala) came asking questions about land and cattle ownership.

Strategies. Support from a trusted official or other influential community member can facilitate trust with individual stakeholders. Sometimes, obtaining the support of a local official, such as a village counselor or chairman, will be sufficient, and local officials who become interested in the project will sometimes even advocate for their neighbors' participation. Other times, it may be necessary to obtain higher-level support, for example from a regional office, in order to obtain the support of local officials.

Although we always first approached village chairmen before conducting interviews, obtaining support from local chairmen was difficult and not always sufficient to allay participants' concerns due to the tensions in Sembabule. In some instances, near-mobs formed, police were called, and one enumerator was even physically attacked. (Fortunately, no one was hurt or arrested.) We quickly pulled our team from Sembabule, directing them to another study region until we could obtain a letter of support from the district chairman.

When the survey team returned with letter in hand, local stakeholders were much less distrustful, and data collection concluded without further incident.

6. Conclusion

In agricultural extension, and throughout the international development sector, there is a strong need for greater understanding of which programs and techniques best achieve desired ends for local populations. While far from being a "silver bullet" for programmatic decision-making, RCTs can provide information about the size of program effects, which can aid decisions about which programs to scale up or expand to make most-efficient use of available resources. Implementing RCTs in agricultural extension programs and other development efforts entails challenges, but experience shows that these challenges are not necessarily insurmountable. A key theme in the strategies we have presented has been communication and stakeholder buy-in—it is crucial for all involved in the project, particularly those with decision-making or implementation responsibilities, to understand and adhere to the study design in order to produce valid results. Likewise, fluid communication makes available stakeholders' deep knowledge about the program and context, which can facilitate evaluation design and mitigate potential local challenges. Carefully and thoughtfully executed, RCTs can yield much more than estimates of causal effects—they provide an opportunity to take a closer look at all aspects of program implementation, generating insights about how programs function that can be used to improve program design or translate beneficial programs to new contexts. Increased investment in rigorous empirical examination of agricultural extension programs, including more expansive judicious use of RCT designs, is the first step toward helping more farmers rise out of poverty.

References

Adhikari, N. K. J., Burns, K. E. A., Eggert, C. H., Briel, M., Lacchetti, C., Leung, T. W., Darling, E., et al. (2005). Randomized trials stopped early for benefit: A systematic review. *Journal of the American Medical Association*, *294*(17), 2203–2209. doi:10.1001/jama.294.17.2203

American Evaluation Association. (2011). *Public statement on cultural competence in evaluation*. Retrieved from http://www.eval.org/aea.culutrally.competent.evaluation.statement.pdf

Bassler, D., Montori, V. M., Briel, M., Glasziou, P., & Guyatt, G. (2008). Early stopping of randomized clinical trials for overt efficacy is problematic. *Journal of Clinical Epidemiology*, *61*(3), 241–246. doi:10.1016/j.jclinepi.2007.07.016

Benstead, L. J. (2010). Effects of interviewer gender and hijab on gender-related survey responses: Findings from a nationally-representative field experiment in Morocco. Retrieved from http://www.polmeth.wustl.edu/media/Poster/BensteadEffectsInter.pdf

Bickman, L. (1985). Randomized field experiments in education: Implementation lessons. *New Directions for Program Evaluation*, *28*, 39–53. doi:10.1002/ev.1408

Bickman, L., & Reich, S. M. (2009). Randomized controlled trials: A gold standard with feet of clay? In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 51–77). SAGE. doi:10.4135/9781412995634.d10

Boruch, R. F., & Wothke, W. (1985). Seven kinds of randomization plans for designing field experiments. *New Directions for Program Evaluation*, *28*, 95–113. doi:10.1002/ev.1413

Buck, S., & Alwang, J. (2011). Agricultural extension, trust, and learning: Results from economic experiments in Ecuador. *Agricultural Economics*, *42*(6), 685–699. doi:10.1111/j.1574-0862.2011.00547.x

Campbell, D. T. (1984). Can we be scientific in applied social science? In R. F. Conner, D. G. Altman, & C. Jackson (Eds.), *Evaluation Studies Review Annual* (Vol. 9, pp. 26–48). SAGE.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*(2), 199–218. doi:10.1177/1098214010366173

Conner, R. F. (1980). Ethical issues in the use of control groups. *New Directions for Program Evaluation*, *7*, 63–75. doi:10.1002/ev.1253

Conrad, K. J., & Conrad, K. M. (1994). Reassessing validity threats in experiments: Focus on construct validity. *New Directions for Program Evaluation*, *63*, 5–25. doi:10.1002/ev.1680

Cook, T. D. (2000). The false choice between theory-based evaluation and experimentation. *New Directions for Evaluation*, *87*, 27–34. doi:10.1002/ev.1179

Dennis, M. L. (1990). Assessing the validity of randomized field experiments: An example from drug abuse treatment research. *Evaluation Review*, *14*(4), 347–373. doi:10.1177/0193841X9001400402

Dennis, M. L., & Boruch, R. F. (1994). Improving the quality of randomized field experiments: Tricks of the trade. *New Directions for Program Evaluation*, *63*, 87–101. doi:10.1002/ev.1687

Duflo, E., Glennerster, R., & Kremer, M. (2007). *Using randomization in development economics research: A toolkit* (No. 6059). Retrieved from http://www.aniket.co.uk/teaching/devt2009/duflo2006.pdf

Duflo, E., Kremer, M., & Robinson, J. (2010). Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya. Retrieved from http://economics.mit.edu/files/6170

Greevy, R., Lu, B., Silber, J. H., & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, *5*(2), 263–275. doi:10.1093/biostatistics/5.2.263

Henry, G. T. (2009). When getting it right matters: The case for high-quality policy and program impact evaluations. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 32–51). SAGE. doi:10.4135/9781412995634.d9

Huddy, L., Billig, J., Bracciodieta, J., Moynihan, P. J., & Pugliani, P. (1997). The effect of interviewer gender on the survey response. *Political Behavior*, *19*(3), 197–220. doi:10.1023/A:1024882714254

Miller, K., Zulu, E. M., & Watkins, S. C. (2001). Husband–wife survey responses in Malawi. *Studies in Family Planning*, *32*(2), 161–174. doi:10.1111/j.1728-4465.2001.00161.x

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315–340. doi:10.1177/109821400302400303

Newman, J., Rawlings, L., & Gertler, P. (1994). Using randomized control designs in evaluating social sector programs in developing countries. *World Bank Research Observer*, *9*(2), 181–201. doi:10.1093/wbro/9.2.181

Patton, M. Q. (2008). sup wit eval ext? *New Directions for Evaluation*, *120*, 101–115. doi:10.1002/ev.279

Raab, G. M., & Butcher, I. (2005). Randomization inference for balanced cluster-randomized trials. *Clinical Trials*, *2*(2), 130–140. doi:10.1191/1740774505cn075oa

Rader, K. T. (2011). Randomization tests and inference with grouped data. *New Faces in Political Methodology Conference, April 30, 2011*. Retrieved from http://qssi.psu.edu/files/NF4Rader.pdf

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. doi:10.3102/0162373707299460

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, *17*(3), 286–304. doi:10.1214/ss/1042727942

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer. doi:10.1007/978-1-4419-1213-8

SenGupta, S., Hopson, R., & Thompson-Robinson, M. (2004). Cultural competence in evaluation: An overview. *New Directions for Evaluation*, *102*, 5–19. doi:10.1002/ev.112

Stevens, S. J. (1994). Common implementation issues in three large-scale social experiments. *New Directions for Program Evaluation*, *63*, 45–53. doi:10.1002/ev.1683

Stringer, A. P., Bell, C. E., Christley, R. M., Gebreab, F., Tefera, G., Reed, K., Trawford, A., et al. (2011). A cluster-randomised controlled trial to compare the effectiveness of different knowledge-transfer interventions for rural working equid users in Ethiopia. *Preventive Veterinary Medicine*, *100*(2), 90–99. doi:10.1016/j.prevetmed.2011.02.001

Van Campenhout, B. (2012). *Mobile apps to deliver extension to remote areas: Preliminary results from Mnt Elgon area*. Retrieved from http://www.grameenfoundation.applab.org/uploads/frontend/mcfile/Blog/Differences_in_Differences_Study_Report_-_Final_July_5_2012.zip

Winters, P., Maffioli, A., & Salazar, L. (2011). Introduction to the special feature: Evaluating the impact of agricultural projects in developing countries. *Journal of Agricultural Economics*, *62*(2), 393–402. doi:10.1111/j.1477-9552.2011.00296.x