A review of five GIF investments focused on Randomised Control Trials

January 2024

Background[1]

The generation and use of evidence underpins GIF's model, and randomized controlled trials (RCTs) are baked into its founding documents. GIF's bylaws define as a critical element "staged financing, with specific caps on funding at each stage, based on rigorous evidence of impact and cost effectiveness for projects that would be scaled with public or philanthropic support." The combination of staging and evidence allows for prudent investment in innovations, which are inherently risky. The more evidence, the more money can be invested. (See Table 1.) GIF promotes innovations' growth to scale by generating further evidence. Innovations which fail to show evidence can be abandoned.

Table 1: GIF's investment stages and their evidence requirements

Stage and funding level	**Purpose and *evidence requirements* for projects with a public sector path to scale (from Byelaws, emphasis added)**
1.Proof of concept ("Pilot") Up to $230,000	To support initial research and design, small-scale pilots and field testing, and initial focus groups or stakeholder consultations to establish viability and user adoption rates.
2. **Testing at an expanded scale and positioning for widespread adoption. ("Test and transition")** $230,000 to $2.3 million	In the case of innovations designed to scale through sustained support with public or private philanthropic funds (or through a hybrid of public and private support), *GIF will provide support for stage 2 applicants to conduct rigorous evaluation of impact and cost effectiveness relative to existing approaches. Evaluations should be designed to isolate the causal impact of the innovation from potential confounding factors… In many cases it will be desirable to collect this evidence through a randomised controlled trial, but GIF will be also be open to other techniques that rigorously measure causal impact (e.g. regression discontinuity designs), where appropriate.*
3.**Transitioning to a widespread scale ("Scale")** >$2.3 million	Stage 3 funding will provide support for transitioning the most successful projects to scale. Stage 3 projects expecting to scale through public sector and/or donor support *should have already demonstrated rigorous evidence of impact and cost-effectiveness in their existing setting.*

[1] Note on authorship: This report is a GIF self-evaluation, in line with standard DFI project evaluation practice. It draws in large part from a commissioned review by Southern Hemisphere, but the findings and views expressed here should be attributed solely to GIF. The principal divergence relates to the potential uses of RCTs. While strongly agreeing with Southern Hemisphere's proposition that the usefulness of RCTs should always be assessed against alternative actions to promote scaling, GIF disagrees with their proposition that *"it is unwise to attempt in the same project to simultaneously create a product or service innovation and test its impact. Similarly, innovations should not still be 'under development' but should already have 'proof of concept' in the Test & Transition stage of GIF funding."* In GIF's view, the wisdom or not of incorporating RCTs into the development of a specific innovation is treated as a topic for evaluation, not a foregone conclusion.

Although RCTs are not the exclusive source of evidence for GIF, they have played a prominent role in its portfolio. These RCTs are always associated with the implementation of an innovation. Sometimes GIF funds the RCT, sometimes the program implementation, and sometimes both.

During GIF's lifetime, the use of RCTs in economic development has massively expanded. At the same time, scholars and practitioners have increasingly questioned the role of RCTs in the scaling process, and RCT practice itself has evolved.

With that background in mind, this report examines the experience of five of GIF's earliest RCT-related investments. As with other completion reports, attention focuses on achievement of outcomes and lessons from implementation. The report adds a cross-cutting dimension on the design and relevance of the associated RCTs.

Innovations and their outcomes

A brief summary of the innovations is given in Box 1. All but Reducing Anaemia (RA) involved behavioural nudges. These deploy incentives or information to overcome constraints on people's ability to improve their health or welfare.

No Lean Season (NLS) was a Scale-stage grant; the others were all Test & Transition stage. Table 2 shows their funding, objectives and outcomes. The goal in all cases was to spur adoption and scale up. For Labelled Remittances (LR) the target adopters were banks. In the other cases, government agencies were envisioned as adopters and scalers-up.

Box 1: Descriptions of the investments

Mobile Conditional Cash Transfers (mCCTs) for Immunisation (2016-2020) GIF funded a 3-year 7-arm RCT in Pakistan to investigate the optimal incentive structure to cost-effectively promote full immunisation coverage for routine childhood immunisation via SMS reminders and small cash transfers to parents.

Grantee: Interactive Research and Development (IRD) Global, in collaboration with Evidence Action and J-PAL

Labelled Remittances (LR) (2017-2019) This project studied the impact of providing Filipino migrant workers in the United Arab Emirates (UAE) with the ability to label the intended use of remittances sent home. The hypothesis was that this would boost their willingness to remit, and that recipients would comply with the labelled intentions, using the funds more productively. The participants in this study were Filipino workers living in Dubai and Sharjah in the UAE. The project built on an earlier, successful 'lab-in-the-field' experiment.

Grantee: Innovations for Poverty Action

No Lean Season (NLS) (2017-2019) No Lean Season was a programme that sought to increase food consumption and income of poor rural households by offering small travel subsidies (grants or no-interest loans) to low-income agricultural workers, enabling them to migrate to find work during the 'lean season' of the year when the demand for agricultural labour falls. Prior, smaller-scale RCTs over 2008-2014 showed that migrants' families increased food intake by 550-700 calories per person per day, and that induced seasonal migration continued in

subsequent years, even after removal of the incentive.[2] GIF's funding was to support rigorous testing at scale in Bangladesh, piloting and testing of the programme in Indonesia, and development of a global strategy for scaling and funding.

Grantee: Evidence Action

Reducing Anemia (RA) (2016-2020) This project aimed to address anaemia and micro-nutrient deficiencies in the Indian state of Tamil Nadu. The intervention involved replacing conventional rice provided free of charge through the Public Distribution Systems (PDS) with fortified enriched with iron, zinc, vitamin A, folic acid and B. The PDS targets poor households.

Grantee: J-PAL South Asia at the Institute for Fiscal Management and Research (IFMR)

Young 1ove (Y1) (2016-2017) Young 1ove sought to test the impact of an innovative education program ('relative risk') aimed at empowering young women in Botswana to reduce their risks of HIV infection and unintended pregnancies. The program was modelled after a Kenyan RCT[3]. In that study, adolescents were given a one hour presentation on the higher relative risk of HIV infection via sex with older men ("sugar daddies') vs younger men. The Kenyan study had found a 28% reduction in girls' pregnancy rate. The Botswana replication tested two modes of instruction: via peers and via teachers. GIF funded the implementation, not the RCT, which was supported by complementary partners, such as J-PAL and the MAC AIDS Foundation.

Grantees: Evidence Action and Young 1ove (separate grant agreements). The RCT was not funded by GIF and was carried out by a collaboration including the Abdul Latif Jameel Poverty Action Lab (J-PAL), Young 1ove, and the Botswana Baylor Children's Clinical Centre of Excellence.

Table 2: Investment funding, objectives, and outcomes

Innovation **Funding agreed and *disbursed***	**Grant objectives**	**Outcomes and progress to scale**
CCTs for Immunisation (mCCTs) $ 856,215 committed	Publish robust data on the optimal amount, schedule and design of conditional cash transfers (CCTs) that would	Outcomes achieved, scale-up in progress. Study[4] was published in a prominent journal, showing cost-effectiveness of small incentives and showing that

[2] Gharad Bryan, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak, 'Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh', *Econometrica* 82, no. 5 (2014): 1671–1748, https://doi.org/10.3982/ECTA10489.

[3] Pascaline Dupas, 'Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya', *American Economic Journal: Applied Economics* 3, no. 1 (1 January 2011): 1–34, https://doi.org/10.1257/app.3.1.1.

[4] Subhash Chandir et al., 'Small Mobile Conditional Cash Transfers (MCCTs) of Different Amounts, Schedules and Design to Improve Routine Childhood Immunization Coverage and Timeliness of Children Aged 0-23 Months in Pakistan: An Open Label Multi-Arm Randomized Controlled Trial', *eClinicalMedicine* 50 (1 August 2022), https://doi.org/10.1016/j.eclinm.2022.101500.

$856,215 disbursed	result in the most cost-effective increase in immunisation rates. Via operational metrics and lessons learned, provide a solid basis for designing a scaled-up policy and service delivery model.	incentive design is as important as incentive amount. In Oct 2021, Open Philanthropy awarded IRD a $25m grant for further scale up, which was subsequently increased to $27.7m. By Oct 2023, more than 0.9 million children were enrolled via incentives; by 2025, approximately 2 million are expected to benefit.
Labelled Remittances (LR) $ 1,736,707 originally committed *amended to $ 1,596,922 disbursed*	Conduct an RCT in the Philippines to test the impacts of a real-world remittance labelling product, including on remittance behaviour of the migrants, and changes in spending patterns and life outcomes of the recipient households Scoping, design and implementation of a study on the impact of remittance labelling in a population of migrant from the South Pacific Islands. In 2020, objective 2 was replaced by a new one, "quantifying the role of international migrant remittances in helping households cope with the pandemic's economic consequences and investigate how these responses are affected by the remittance labeling innovation." (in the Philippines)	Outcome partially achieved, evidence does not support scale up. A working paper version of the study for objective 1 was published[5]. The RCT did not find an increase in remittances, on average. The labels did not increase recipients' expenditure, or shift the expenditure toward the labelled uses. Preliminary findings for objective 3 were reported but a paper has not been published (as of Sept 2023). An unanticipated methodological paper was published[6]. Anonymized datasets for objectives 1 and 3 were published[7].
No Lean Season (NLS) $ 2.571,210 committed.	In Bangladesh, at least 30,000 travel subsidies disbursed and develop the capacity to disburse at least 90,000 travel subsidies if the results of the RCT are positive	Outcomes partially achieved, no progress to scale More than 130,000 loans were disbursed. 2017 preliminary impact evaluation did not show positive results. The lack of impact

[5] Giuseppe De Arcangelis and Dean Yang, 'A Field Experiment among Filipino Migrant Workers in the UAE', RSIE Discussion Papers (University of Michigan, February 2022). https://fordschool.umich.edu/rsie/workingpapers/Papers676-700/r684.pdf

[6] Giuseppe De Arcangelis et al., 'Measuring Remittances', *Journal of Development Economics* 161 (1 March 2023): 103004, https://doi.org/10.1016/j.jdeveco.2022.103004.

[7] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PCMAWS

$ 2,461,901 disbursed of which $461,636 was returned to GIF	RCT results in policy-relevant, robust evidence In Indonesia, an operational model with 1000 subsidies disbursed and successful RCT A global strategy for program development including funding approach	may reflect implementation shortcomings. Hence the results were not robust. Evidence Impact discontinued the study due to lack of evidence of impact and concerns about the implementing organisation. Indonesia operations were not carried out.
Reducing Anemia (RA) $ 1,303,611 committed *$ 531,752 disbursed*	Conduct the RCT, including rice fortification and delivery; disseminate findings to policymakers, and publish the study.	Outcomes not achieved, no attributable progress to scale. The study was discontinued due to implementation problems, including regulatory changes, problems in manufacturing and distributing the fortified rice, inconsistent testing, and COVID lockdowns. However, the government independently instituted a similar pilot program starting in 2020.
Young 1ove (Y1) $ 690,238 committed *$ 363,425 disbursed*	Support evidence-based efforts to scale the 'relative risk approach' via: Young 1ove: by convening key stakeholders regularly, piloting in at least one new country, updating curriculum based on the evidence, and via organizational development. Evidence Action: by dissemination of results and strategic inputs to Young 1ove	Grant outcomes partially achieved. Post-grant, continued testing, adaptation, and program expansion followed. Preliminary results showed that pregnancy rates were lower than assumed; that 'sugar daddies' were younger than assumed; and found that the impact on pregnancy was ambiguous. Following this, a second year of support was not provided by GIF. Subsequently, a published paper reported that the intervention reduced the risk of pregnancy. Young 1ove (now Youth Impact) has continued to analyse the data and has modified the program based on the evidence, for instance leveraging national youth service programs to deliver the intervention. The organisation has also grown in size and programmatic offerings.

Four of the five investees completed their RCT, an immediate goal of GIF's grants.

- The mCCTs study supported cost-effectiveness of their incentive, and provided actionable insight into incentive design.
- LR's study was not supportive of the impact of labelling.
- Y1's initial results were ambiguous on impact and questioned initial assumptions about pregnancy rates and the age of sugar daddies. A subsequent published analysis showed that

the program reduced teenage pregnancy. Further work has focused on improving efficacy by choosing the right channel of delivery.
- NLS's 2017 RCT did not support effectiveness of the program. Subsequent analysis suggested that this may have been due to problems in implementation.

Two of the innovations have proceeded toward scale, the ultimate goal of GIF support.

- mCCTs has attracted additional funding and is being implemented in cooperation with the Pakistan government at larger scale.
- Y1 modified and expanded its program on sexual health; organizationally it has branched out into education.

Implementation experience

Additionality

Interviews (by the commissioned evaluators) with grantees mentioned a number of ways in which GIF support adds value to their projects, including willingness to fund the innovation as well as the RCT, networking and introductions to stakeholders (e.g., in RA it facilitated introductions with government stakeholders) and GIF's reputation for good investment decisions assists with 'crowding in' other funders. As one respondent said, "Without their support, none of this would have happened." In four of the five investment records, GIF sees its additionality as its commitment to evidence and scale; it also mentions willingness to fund regions or topics that have not received sufficient attention (thereby being able to produce evidence to attract other funders).

Diligence and investment process

In the contracting phase, GIF conducted due diligence and risk analysis, and accepted a residual level of risk in line with its mandate. GIF carefully considered ESG issues, undertaking mitigatory actions where needed. All RCTs received ethical approval by Independent Research Boards, though this was not always documented in GIF memos. For NLS, the reviewers and the decision panellists expressed concerns about negative social implications for the health and security of the migrants and the well-being of the families left behind, and recommended that the social standard risk needed more careful consideration. The grant established a risk register that Evidence Action would report on. Evidence Action had already identified a risk of inducing underage migration and established procedures to mitigate that risk.

Operational phase and supervision

In all five cases, implementation deviated from the original plan.

- **mCCTs.** The research protocol was amended to accommodate a new law requiring identification of money transfer recipients.
- **LR.** The original plan was for the implementing bank to include the 'label' (intended use of funds) as part of the transfer documentation. The bank withdrew and its replacement did not agree to modify the transfer process. The project introduced a companion smartphone app to transmit the 'label'.
- **NLS.** The 2017 round of the RCT did not find the hoped-for impact on migration. Evidence Action surmised that this was due to problems in implementation. They planned operational research to address the issue, to be followed by another RCT evaluation. In 2019 there was a tragic accident in which four underage migrants died. Later that year, Evidence Action learned of bribery allegations against the local implementing organisation. Evidence Action

decided to terminate the program rather than seek a new partner[8]. GIF's grant was cancelled and unspent funds were returned.

- **RA.** RA was a more complex intervention than the others reviewed here. It was known that iron supplements benefit anaemic people. What was to be tested is whether a program could reliably manufacture and deliver appropriately fortified rice to the intended beneficiaries to get that end result. Manufacturing and distribution proved problematic. New food safety regulations led to a delay. The researchers had difficulty finding a reliable manufacturer. Initial testing of the fortified rice yielded inconsistent results, creating a delay that was then exacerbated by the COVID crisis. The grant was cancelled.
- **Y1.** Initial RCT analyses found mixed impact on reducing pregnancy. The investment record had planned for a second year of support. This included an expectation to continue funding regardless of the RCT findings, contingent on approval of revised goals. In the event, however, GIF cancelled the grant.

The commissioned review found that GIF has a relatively good trail of monitoring data such as narrative and financial reports. However, staff reactions to these reports were not well documented, potentially impairing organisational learning. Staff monitoring was supplemented in three cases by advisory committees. At the time of these grants GIF did not have a strong close-out process. Grantee final reports were often weak in documenting outcomes and making overall assessments.

Relevance of design and implementation for scale

Were the RCTs well-designed, from a technical viewpoint?

Third party reviews were conducted by external reviewers for all the studies under review, but only two did an assessment of the RCT methodology itself. The others approved of the methodology as a whole or approved because of their confidence in the team members proposing the RCT.

Was an RCT the right approach, at the time, to inform the innovation's journey to scale?

Behavioural nudges are well-suited to RCTs. The nudges can be offered to a randomly selected treatment group. Outcomes are compared to those in a control group. This allows researchers and policymakers to test whether the nudge really had an impact. Without an RCT, one might argue that the causality worked in reverse: people tending toward better outcomes are more likely to accept the nudge. For instance, girls interested in signing up for a training course on HIV prevention might be at lower risk than others.

However, none of the proposals considered any alternatives to an RCT or justified why an RCT was the best methodology for the study, given the proposed pathways to scale or evidence users. The commissioned review reported that 'Most of the PIs we interviewed said that an RCT was not the only methodology they could have proposed to evaluate the impact of the study that they were conducting.'

A key issue is whether an innovation is ripe for an RCT. During scale-up, innovations face multiple challenges :

- Testing operational mechanics – do the logistics work in practice? Are operational plans faithfully carried out by field workers?
- Verifying causal impact – is the innovation really making a difference?
- Gauging cost-effectiveness – does it look as if it will be cost-effective at scale?

These issues are tested repeatedly, often concurrently, as an innovation scales up from hundreds of people to millions. In the process, the innovation may evolve radically. During that process, an RCT

8 https://www.evidenceaction.org/were-shutting-down-no-lean-season-our-seasonal-migration-program-heres-why/

might be essential for generating confidence for the next stage of scale, or for evaluating design alternatives. On the other hand, working out logistical kinks might be a necessary step before proceeding to a test.

The record here is mixed: two RCTs that were clearly the right approach, two that might have benefited from better advance planning, and one that foundered on logistical challenges.

- For mCCTs, implementation built on prior experience but still was able to adapt in progress to changes in the regulatory environment. The multi-arm RCT design provided valuable information on the cost-effectiveness of different designs and levels of incentive, applicable to the next stage of scale up. GiveWell, a donor and advisor to philanthropists, had supported the RCT. After its completion, GiveWell's recommendation was important to mCCTs' receipt of a major grant. The RCT results also led to the government's decision to scale up implementation.
- For LR, the switch to an app-based label represented a modification to the original design, but the RCT nonetheless contributed valuable information on the effectiveness of a labelled remittance.
- NLS built on a sequence of prior RCTs, incrementally scaling up to larger populations based on prior learning. A retrospective on the 2017 RCT reported that "subsidies mainly reached those who would have migrated anyway, and the programme was promptly discontinued" and that discontinuation was justified.[9] As noted above, however, Evidence Action hypothesized that that potentially-correctable implementation issues were responsible for the outcome.
- Y1, in retrospect, might have collected more descriptive data up front or conducted multiple A/B tests, in order to test assumptions and modify the program further and sooner. At the same time, the RCT did ultimately provide valuable lessons Y1 could then act on to further refine the program, and the organization has incorporated rapid A/B testing in implementation and scale up.
- RA did undertake pre-testing of the fortified rice's acceptability to consumers, but was unsuccessful in working out manufacturing and distribution, leading to cancellation of the RCT.

Did the grants position the investees to inform and influence scale-up or replication of the innovation?

The two innovations (Y1 and mCCTs) that proceeded to scale benefited from strong partnerships with the scaling government. The grantees combined research and implementation experience. mCCTs is a particularly good example of planning for scale. IRD, the researcher/implementer, had a long-standing relationship with local immunisation authorities and had been involved in setting up an Electronic Immunisation Registry. IRD consulted with other donors and implementors. In contrast, RA began with strong government support, including financial, but this was not sufficient for program completion and scaling.

The other two innovations lacked well-defined scale plans. GIF recognized from the start that securing funding and buy-in for NLS would be challenging. LR lost its initial bank partner. The weak results of the RCT rendered moot the question of scale.

Communications plans are potentially critical to the scaling process. Explicitly or implicitly, the purpose of an RCT is to provide actionable information to decision-makers and stakeholders. Only

[9] Ahmed Mushfiq Mobarak, 'Assessing Social Aid: The Scale-up Process Needs Evidence, Too', *Nature* 609, no. 7929 (September 2022): 892–94, https://doi.org/10.1038/d41586-022-03039-2.

two of the five investees produced communications plans – the two successful scalers, mCCTs and Y1.

Conclusions, recommendations and reflections

It is perilous to generalize 'lessons' from a small sample. GIF has several additional RCT-related investments that are ripe for evaluation, and they will provide a wider evidence base. Nonetheless, the following issues emerge for consideration.

1) Successful innovators are likely to have experience in implementation as well as research, and to have strong ties to scalers-up

This is not news to GIF. But it is worth keeping in mind when assessing the risks to scale, and in deciding how actively GIF wants to play the role of a 'scale entrepreneur' in promoting successfully-performing innovations to governments and funders.

2) Establish the purpose of a proposed RCT, and consider against alternatives

Proposals should be clear on the purpose, relevance, and cost-effectiveness of an RCT. For instance, it is to establish causality, test innovation design, measure cost-effectiveness, or a combination? Is the innovation ripe for an RCT, or is it more appropriate to test operational or design issues first? To what extent is the journey to scale iterative, and how relevant will these results be to the decision to scale or replicate?

3) Ensure that there is a strong communications plan

Communications are key for ensuring that scaling happens; they are necessary throughout the scaling process to keep stakeholders informed and engaged and at the end, to share results to either encourage further scaling if the results are positive or to share scaling lessons from projects that failed to scale. Information also needs to be timely to inform policy and programming decisions.

GIF investments should, during the first year of the investment, produce a communications plan as one of the deliverables tied to a payment tranche. The plan should also clearly indicate who is responsible for implementing the plan. The plan should take cognisance of possible policy windows or funding cycles relevant for the innovation. The communications plan should include KPIs that measure at least reach and ideally impact. The approved budget for the investment should contain an explicit line item for communications. As part of the final report, grants should include a comprehensive report on the communications activities carried out throughout the project and their likely impact. All research outputs should be listed in this report and copies should be provided to GIF.

4) Examine how formative or operational evaluations fit into GIF's Investment Policy

For some innovations, operational issues – such as how to ensure fidelity of implementation – may be critical areas for testing. Yet these may not fit well into the criteria for Test & Transition-stage investments, nor be feasible within the tight budgetary limits of the Pilot stage. Can investments of this type fit into the Investment Policy in a way consistent with the Byelaws?

5) Be flexible

Again it will not be surprising, but worth keeping in mind, that implementation rarely follows the plan. Flexibility to deal with regulatory changes and implementation setbacks is essential. GIF should explicitly consider its appetite to continue support for an innovation where RCTs yield ambiguous or null results. There is a tension between the stage-based investment approach, based on an option to abandon unpromising innovations vs. support for innovators to learn from failure and improve.

6) Risk management

Scaling of complex social innovations is likely to have trade-offs which could negatively affect the communities they are intended to benefit. During the period in which the investments under review were active, GIF did have good systems for due diligence and risk mitigation, yet safeguarding issues arose for NLS. While some degree of risk is unavoidable in an organisation devoted to de-risking, for investments with a high ESG risk, GIF could explore alternatives for enhanced risk mitigation. Direct monitoring of implementation is unlikely to be affordable and feasible within GIF's model.

7) Grant-making and monitoring

Improvements could be made to the grant-making systems at GIF. During the time under review, there were gaps in the reporting and in GIF's responses to reports it receives from grantees. The KPIs used for monitoring scaling were not always adequate.

Potential areas for improvement include:

- Encouraging grantees to create policy advisory committees when appropriate
- Developing standard KPIs that better capture progress toward scaling
- Applying a standard template for grant completion reports such as that of the International Development Research Center[10].

[10] https://idrc.ca/en/guidelines-preparing-final-technical-reports (Accessed: 23 November 2022).