Journal of Development Effectiveness
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjde20

A checklist to avoid pilot failures: lessons from a set of women's economic empowerment initiatives

Sara Johansson de Silva[a], Pierella Paci[b] & Josefina Posadas[b]
[a] World Bank, Malmö, Sweden
[b] World Bank, Washington, DC, USA
Published online: 09 Oct 2014.

To cite this article: Sara Johansson de Silva, Pierella Paci & Josefina Posadas (2015) A checklist to avoid pilot failures: lessons from a set of women's economic empowerment initiatives, Journal of Development Effectiveness, 7:1, 90-110, DOI: 10.1080/19439342.2014.963882

To link to this article: http://dx.doi.org/10.1080/19439342.2014.963882

Journal of Development Effectiveness, 2015
Vol. 7, No. 1, 90–110, http://dx.doi.org/10.1080/19439342.2014.963882

Routledge
Taylor & Francis Group

A checklist to avoid pilot failures: lessons from a set of women's economic empowerment initiatives

Sara Johansson de Silva[a], Pierella Paci[b] and Josefina Posadas[b]*

[a]*World Bank, Malmö, Sweden;* [b]*World Bank, Washington, DC, USA*

Pilot programmes have gained significance in donor-supported development interventions because of the growing emphasis on measuring impact. The Results-based initiatives (RBI) were conceived as pioneering pilots expected to acquire rigorous evidence on effective interventions to foster women's economic empowerment. However, they fell short of providing clear or generalizable conclusions on women's economic empowerment due to design and implementation problems. The RBI nevertheless offer important lessons on common traps in pilot design and implementation. This article synthesises 10 lessons from the RBI as a checklist to avoid pilot failure, intended for practitioners in any area of development.

Keywords: pilots; impact evaluation; entrepreneurship; training; skills; women; economic empowerment

1. Introduction

Pilot programmes are increasingly becoming a feature in experimenting with innovative approaches to development. It is not difficult to see why. When adequately designed, implemented and evaluated, small-scale programmes can provide a rigorous, fast and relatively inexpensive testing ground for larger scale interventions. However, small is not always beautiful: without careful planning, implementation and follow-up, pilots may fail. A pilot that shows programme failure is not automatically a failure: instead, a pilot failure occurs when the programme evaluation cannot provide answers to the question it had set up to address, wasting resources instead of supporting judicious use of development funds.[1] Building on experience from the Results-based initiatives (RBI), this article derives a checklist designed to navigate more effectively those small, but crucial, obstacles and oversights in pilot design and implementation, and in the evaluation/monitoring methodology that may lead to pilot failure and crash landing.

The RBI were a pioneering programme of pilots expected to produce rigorous evidence on effective interventions to foster women's economic empowerment. When conceived by the World Bank and United Nations Women (UN Women), then UNIFEM, in 2007, the programme was innovative in its use of small-scale pilots and rigorous statistical techniques to evaluate the relative effectiveness of different interventions to promote women's economic empowerment in different country settings. The focus on business training interventions as means to reduce gender gaps was also novel since much of the literature and policy dialogue had been centred on gaps in health and education (McKenzie 2010; McKenzie and Woodruff 2012).

Given their innovative character and scale, the RBI had the potential to provide policy makers with invaluable lessons on the potential impact of different policy packages.

*Corresponding author. Email: jposadas@worldbank.org

Unfortunately, ex-post evaluations show that the programme missed that opportunity as the pilot findings were too weak to derive any robust and generalizable conclusions on the effectiveness of alternative interventions. Many lessons on impact therefore remain 'not yet learned'. By contrast, the RBI provide useful lessons on the *dos and don'ts* of pilot design and implementation for development practitioners.

This article builds on these lessons to derive a checklist of necessary conditions that need to be satisfied for a pilot to be meaningful and worthwhile. It draws on a set of background reports on individual pilots and on two summary reports prepared by UN Women (previously UNIFEM) and the International Center for Research on Women (ICRW) (UN Women 2011; Golla 2011).

Many of the points raised are common sense, and none are new. Some, but not all, relate to the specific requirements of rigorous statistical impact evaluations (IEs) in pilots. Nonetheless, pilots continue to fail because these simple points remain systematically neglected in design, implementation and monitoring.

The literature on IE and project monitoring is vast and largely directed to a technical audience due to the rigorous statistical methods required to ascertain statistically significant effects (Duflo, Glennerster, and Kremer 2007). Although there have been recent and valuable attempts to present the technical aspects in a simple manner to reach a wide audience (World Bank 2006; Gertler et al. 2011; Khandker, Koolwal, and Samad 2010), there is a void in presenting some of those lessons using largely nontechnical language as is done in the broader monitoring and evaluation strand of papers (Clark, Sartorius, and Bamberger 2004; Bamberger 2009).[2] The objective of the article is to fill in this void by presenting a checklist to help a broad range of interested actors to avoid future costly pilot failures.

The next section summarises the main design features of the different RBI interventions. Section 3 takes stock of the findings of the individual RBI IEs. Section 4 focuses on the main lessons learned on pilot design, implementation, and monitoring and IE. Section 5 concludes.

2. The nuts and bolts of the RBI pilots

The RBI are comprised of eight country pilots under a common umbrella programme whose objective was to generate robust and consistent evidence that could be generalised. This article focuses on the five pilots that have been implemented and fully evaluated to date.[3] As shown in Table 1, all single interventions had common elements, but also presented variations in design, implementation and IE methodology, reflecting differences in local contexts and priorities. The similarities and dissimilarities span the design of the interventions, implementation mode and IE.

The RBI pilots shared many elements. They were small programmes, with the common objective of fostering women's economic empowerment through enhancement of women's skills, assets and decision-making. They had limited budgets and were expected to be relatively quick to implement and evaluate. Their common focus was on delivering training, and they were all expected to provide measurable and statistically valid evidence on programme impact through rigorous IEs. All IEs, except Liberia, used some type of randomised control trial (RCT). Finally, UNIFEM was the common executing agency, relying on local implementation agencies and with a high degree of involvement of government and local nongovernment organisations.

However, the interventions also differed substantially in the setting in which they operated and in the design. The programme spanned different regions with very different cultural, geographic and socioeconomic situations. It covered low- and middle-income countries with

Table 1. Summary of RBI pilot interventions.

Objectives	Intervention
Egypt: gender equity model	
Improve gender equity in access to jobs, wages, career development, working conditions and employment participation.	Foster good gender equity practices in the private sector by helping firms formulate and achieve gender equity goals through human resource training.
Kenya: strengthening export competitiveness of women beadworkers	
Increase productivity and earnings of women in beadwork by increasing access to export markets.	Provide training and mentoring to improve design, marketing and business skills; help identify and access larger markets.
Liberia: value-added cassava enterprise for the Ganta Concern Women's Group	
Increase economic security, strengthen livelihoods, empower participants and promote the cassava industry as a growth sector and poverty reduction strategy.	(i) Training on improved farming techniques for growing cassava; (ii) activities in the Ganta group to help them negotiate for more land; (iii) grants to pay for tools, purchasing inputs and so forth; (iv) a cassava farina plant and (v) additional activities to strengthen the organisation's governance and gender sensitivity.
Mekong: improving bamboo handicraft value chains for women's economic empowerment	
Increase productivity and earnings of women in the bamboo handicraft production.	(i) Groups training on production of bamboo handicraft products; (ii) study tour for producer groups and traders and (iii) equipment.
Peru: strengthening the economic empowerment of women microentrepreneurs in Lima	
Increase productivity of women microentrepreneurs and their bargaining power in household.	Provide training in business practices, marketing and life skills to women microentrepreneurs with land title (and thus collateral for credit).

Source: Authors' compilation.

different levels of institutional capacity; rural and urban settings; poor and nonpoor groups. In Egypt, Kenya and Peru, they delivered only training, while in Liberia and Mekong training was complemented with grants and equipment. The Egypt pilot aimed at reducing gender wage gaps amongst wage workers, while the others focused on increasing the earning capacity of self-employed women. In Kenya and Peru, training was limited to business skills development, while in Liberia and Mekong this was combined with training on production techniques. The interventions targeted producer groups in Kenya, Liberia and Mekong, but individual entrepreneurs in Peru. They varied also in the technical details of the randomisation roll-out, the final methodology of the IE, the capacity of the implementing agencies and the degree of government involvement. For example, in Peru the evaluation team was experienced and had previously worked successfully with the implementing agency, and in Egypt the intervention was previously tested in other countries.

The considerable variation in design, methodology and implementation challenges across the different pilots introduced too many degrees of variation to allow generalizable conclusions to be drawn.

3. Lessons not yet learned on impact

The impact of the RBI pilots was assessed on three dimensions of women's economic empowerment (World Bank 2011): (i) economic opportunities, (ii) endowments – the

stock of human capital and assets and (iii) agency – the ability to make effective choices leading to desired changes. The questions addressed were: (i) *what impact* did each pilot have on these dimensions; (ii) *how reliable* are the findings – based on statistical significance/robustness of results and validity of the IE design and (iii) do *consistent messages* emerge across countries?

Findings from the individual IEs on each dimension are presented in Tables 2, 3 and 4 respectively.[4] The main emerging message is the lack of consistent results: the observed impacts were mixed and estimates were not always reliable.

3.1. *Economic opportunities*

Although promoting economic opportunities was a major objective of the programme, no consistent and statistically significant impact on labour market outcomes was found. In Egypt, there was no increase in hiring of women, in Liberia the cassava farina plant did not generate sustainable jobs and in the other pilots the number of women engaged in targeted activities remained constant. Table 2 also shows that the impact on earnings was mostly weak. In Egypt, the impact on women's wages and the gender wage gap was statistically insignificant and so was that on revenues of self-employed women in Kenya, Liberia and Mekong. Only in Peru, the intervention led to a statistically significant improvement in business earnings.

Nonetheless, some encouraging findings emerged. In Peru, bookkeeping was introduced, business practices improved and sales increased by nearly 20 per cent when general training was combined with technical assistance in business development. In Mekong, product quality increased and new designs were introduced. In Egypt, male workers displayed stronger awareness of existing gender inequalities, of supervisors' fairness and of gender equality in career development and training opportunities.

3.2. *Endowments*

The findings on the impact of the interventions on women's labour market skills were more positive. Women in treated Egyptian firms saw their on-the-job training increased. Liberian focus group participants pointed to training in literacy and farming practices as the most useful part of the intervention. However, not all training translated into practice. In Liberia, the suggestion by the trainers to buy in bulk to benefit from lower prices was never adopted. In the Mekong Valley, the splitting machines offered by the project were hardly used because they produced bamboo strips that were too large for small handicrafts.

By contrast, where evaluated – Mekong and Peru – there appeared to be no statistically significant increase in investment in children's education or health, except for a positive impact on boys' education in Lao PDR.

3.3. *Agency*

The interventions showed a mildly positive impact on agency through participation in networks and associations. Participants in Cambodia and Peru reported being more active in decision-making over large household expenses after the intervention, but this was not the case in Lao PDR. In the Mekong Valley, women moved from home-based to group production of bamboo handicrafts to better learn from each other. In Peru, more women joined business associations and informal lending networks.

Table 2. Impacts on economic opportunities.

	Women's labour force participation	Technology of production	Wages and sales	Career	Statistically reliable results?
Arab Republic of Egypt	Absenteeism increased for both men and women in the treated firms. However, this result can be driven by seasonality effects.	Not applicable.	No effects on gender wage gap.	No effects at firm level on hiring, training, promotion or representation of women in management.	Yes With caution due to (i) heavy reliance on self-reported information and (ii) no robust statistical specification.
Kenya	No new workers attracted to bead handicrafts. Beneficiaries initiated other income-generating activities.	Groups did not embrace suggestion to purchase in bulk to lower input price. Introduction of new designs had no impact.	No effect on sales strategies or traders' performance at group level. No effect on revenues of beneficiaries in either bead-only work or in groups that combined beadwork with other activities.	Not applicable.	No Weak IE design. Weak implementation: curricula of training.
Liberia	Very few paid jobs created, and these jobs were only available to those living near Ganta Concern Women's Group.	Class training on farming practices had no effect. On-site field demonstrations were effective. Plant operates at less than 10 per cent of capacity.	Increased access to land and labour, and improved farming techniques, leading to higher cassava production in group farms but with high variability of impacts across groups. Caveat: the cassava was sold only to the processing plant which could not process it and many fields remain unharnessed.	Not applicable.	No Weak IE design: no valid control group. Weak intervention design: underestimated constraints such as literacy.

Mekong	No change in the probability of households producing BHP.	Increase in quality and new designs. Bamboo-splitting machines not used. No effect on access to raw bamboo.	No effect on the share of bamboo activity in household income. Sales increased during the low season, but only in Cambodia.	Not applicable.	No Weak IE design: small sample.
Peru	A few beneficiaries closed their businesses, possibly because they realised the businesses were not viable, but may also reflect the normal high exit rate of small firms.	Women who received GT + TA introduced a few good business practices (bookkeeping and executing innovations) and increased reliance on formal credit; use of informal credit increased for all beneficiaries.	Sales increased by 19 per cent among women that received GT + TA. In a second round, results remained positive, but their size slightly decreased and their variation increased.	Not applicable.	Yes Robust design and implementation + credible IE.

Note: BHP = bamboo handicraft products; HH = household; GT = general training; GT + TA = general training and technical assistance.
Source: Authors' compilation based on Golla and Jones-Demen (2011); Golla and Saggers (2011); Golla and Selim (2011); Knowles (2011) and Valdivia (2011a, 2011b).

Table 3. Impacts on endowments.

	Women's labour market skills	Children's health, nutrition, and education	Assets	Statistically reliable results?
Arab Republic of Egypt	Women in treated firms received on-the-job training, and their perceptions of their labour market skills improved.	Not measured.	Not measured.	Yes Based on self-reporting by workers.
Kenya	Training in new designs, but no changes were observed in the beads produced. Women valued exposure to trade fairs/market places.	Not measured.	Reduction in group's savings and material assets.	No Weaknesses in design and implementation.
Liberia	Mixed reactions to training for agricultural practices; most women were not interested in classroom training, but did value the literacy training.	Not measured.	Not measured.	No Weaknesses in design and implementation.
Mekong	Training of the trainers succeeded, with 1354 group participants, most of who were women, who perceived the training as useful.	Increase in spending on boys' education in Lao PDR; decreased spending on personal care items for men; no effect on child labour.	Not measured.	Yes With caution, based on self-reporting by producers.
Peru	Increase in women's time devoted to studies for those that received both training and targeted assistance; decreased for those who received only training.	Decrease in time other females in the household spent in business and other work, for beneficiaries of both training and targeted assistance; no effect on time spent on studies by children (7–13 years old).	Increase in the use of informal credit (though there is some crowding out of formal credit).	Yes Robust design and implementation, credible IE.

Source: Authors' compilation based on Golla and Jones-Demen (2011); Golla and Saggers (2011); Golla and Selim (2011); Knowles (2011) and Valdivia (2011a, 2011b).

Table 4. Impacts on agency.

	Household decision process	Groups and networks	Perceptions	Statistically reliable results?
Arab Republic of Egypt	Not applicable.	Not applicable.	Employee satisfaction increased for both men and women. Awareness of gender equality increased for both men and women. Mixed results on workplace: discrimination increased (women only) and experiences of a hostile work environment decreased.	Yes With caution, many outcomes rely on self-reported information. Weak design: small sample.
Kenya	Not measured.	No change in the number of groups registered or the average number of group members.	Young women with some education perceived the project as beneficial. Market visits and trade fairs found useful to learn about marketing strategies.	No Weak IE design: small sample.
Liberia	Some women reported feeling more confident and having higher self-esteem. Families' perceptions about group's work were mixed due to extensive time commitment of beneficiaries with little economic return.	Membership in group did not change. No ability to manage the plant independently. No sense of ownership towards the enterprise, and no sense of responsibility. Diagnostic of the organisational problems did not engage the group in seeking changes. Leadership training, valued by individuals, did not affect group governance or transparency.	Overall sense of dissatisfaction with the intervention. Group leaders would have preferred guidance with the management of Ganta and not the building of the farina plant. Some satisfaction with training in farming techniques.	No Based mostly on qualitative data from focus groups.

(*Continued*)

Table 4. (*Continued*)

	Household decision process	Groups and networks	Perceptions	Statistically reliable results?
Mekong	Significant positive effects on likelihood of bulk purchases (Cambodia only). No significant effects on gender roles in BHP-related decisions. No effect on gender roles within the HH.	Increased number of households working with a producer group.	Perceived constraints were access to inputs (too expensive), access to traders and low product prices.	No Weak design: lack of statistical power to control for clusters.
Peru	Increased women's participation in HH decision process for those that receive both GT + TA. Larger effects among young, single, less educated beneficiaries with land title and medium dependence on business.	Increased participation in business associations for all beneficiaries.	Increased acceptance of working women. No effect on perception of gender roles within the HH.	Yes Robust design and implementation and credible IE.

Note: BHP = bamboo handicraft product; Ganta = Ganta Concern Women's Group; GT + TA = general training and technical assistance; HH = household.
Source: Authors' compilation based on Golla and Jones-Demen (2011); Golla and Saggers (2011); Golla and Selim (2011); Golla (2011); Knowles (2011) and Valdivia (2011a, 2011b).

4. Lessons learned on pilot design and impact evaluation

The lack of robust and generalizable results on the impacts of the programme is disappointing. Worse is the fact that it is impossible to ascertain the extent to which the weak findings reflect the inherent ineffectiveness of the interventions – intervention failure – as opposed to weaknesses in the design of individual pilots and the umbrella programme, their implementation and evaluation – pilot failure. In this respect, the RBI represent a costly missed opportunity to generate valuable evidence on effective levers to improve women's economic environment.

However, the shortcoming of this programme can provide a number of lessons on how to avoid basic errors that may limit the potential of pilots from the outset. This section focuses on lessons on pilot design and IE. A key conclusion is that if time and money are in short supply, expectations need to be managed. Pilots that do not incorporate rigorous IEs cannot provide robust knowledge. If the time, resources and capacity to undertake rigorous IEs cannot be found, a more economical evaluation to test the particular programme should be considered, based on good-quality quantitative and qualitative data.

Table A1 (in the Appendix) shows that the RBI complied with several of the established principles of good practice for IEs. The pilots were justifiable, given the existing knowledge gaps in the area, and they all complied with ethic norms (although not formally approved by ethics committees). Rich – including qualitative – data were collected at the level of the expected impact, and beyond, to capture changes in group dynamics and social norms. Innovative performance indicators were used to monitor difficult-to-measure outcomes and progress relatively unexplored areas – such as household decision-making and group dynamics.

Despite some problems with selection and randomisation,[5] all RBI – except Liberia – had valid control groups reflecting the chosen IE methodology. However, compromises had to be made in several areas, including the selection of the control group and the recruitment of beneficiaries. In Kenya and Mekong, control villages had to be chosen in different provinces – where inhabitants specialised in different crafts – following severe resentment from individuals in neighbouring control villages. Moreover, in 11 of the 43 Mekong villages, the treatment/control decision was based on political consideration rather than randomisation.[6] In Egypt, the initial lack of participating firms forced the team to replace the RCT with a quasi-experimental design. Combined with the other weaknesses in design and implementation described in Sections 4.2–4.5 these problems made it difficult to derive conclusive results, leading to the following lessons.

4.1. Lesson 1: be realistic about the learning component

Balancing ambition with financial resources is a major challenge for pilots. The monitoring, evaluation and learning components are essential parts of all pilot programmes, as they ensure their replicability and scaling-up. However, teams should be aware that the learning aspect of pilots can add considerable fixed costs and time requirement, beyond those of small programme interventions with the same number of beneficiaries. This is due to the higher costs of (i) due diligence and preparation, including of IE; (ii) need for a minimum size (lesson 2), (iii) detailed monitoring and evaluation and (iv) dissemination. It is essential to take this into account in the planning of the intervention and budget accordingly.

The budget envelope for each individual RBI – except for Mekong Valley – was approximately US$600,000, equally divided between implementation and IE.[7] The

programme design – involving building value chains and training in highly complex, low-capacity environments and full-scale IEs – was too ambitious for the funding. The fact that resources were not explicitly allocated to monitoring or dissemination further limited the capacity to learn from its implementation.

4.2. Lesson 2: small is not always beautiful

Restrictive rules on who can participate are often used to lower the cost of pilot implementation – by lowering transaction costs and making coordination/communication easier. But these rules risk leading to sample sizes that are too small to derive statistically significant conclusions. For the pilot to give results that are meaningful for policy making, the number of eligible beneficiaries has to be large enough for the conclusions to be relevant to the whole population of (potential) beneficiaries (external validity), and it needs to be consistent with the evaluation technique chosen to ensure reliability of results (internal validity).

The RBI pilots had a relatively small number of direct beneficiaries – from 250 in Liberia to almost 1500 in Peru – selected according to geographical or socioeconomic criteria. These small numbers limited their power to give robust results as the participants proved to be too few. Power calculations, of the instrument to determine the sample size needed to provide statistically significant results, were only done for two of the five pilots. For Peru, the sample was sufficiently large; however, this was not the case in the Mekong pilot where a decrease in the number of clusters – from 20 to 18 – and the number of observations per cluster indicated the lack of power.[8] The pilot continued in spite of the predictions about power, suggesting the lack of exit clauses in the set-up of the pilots.

4.3. Lesson 3: patience is a virtue

A short time frame is typical of pilot programmes and in line with the overarching objective of generating evidence quickly, initiating the policy dialogue promptly, triggering further research, and/or scaling up the intervention. However, a compressed time frame risks cutting down essential stages of the process and reducing the attention given to pilot development, including client consultations. It can also shorten the time devoted to understanding the theory behind the intervention: its expected impacts, the mechanisms through which the impacts will be achieved and the potential distributional effects. Ultimately, it may result in delays along the way due to a higher number of unforeseeable adjustments.

The RBI were expected to show quick results – within 1–2 years of project initiation. Overall, this time frame was too ambitious for programmes that involved complex change processes, including improvements along the entire value chain (Liberia, Kenya, Mekong), organisational development, changed preferences and attitudes in the private/public spheres, improvement in business practices and changed relations at household level. Achieving such complex transformations is a difficult and time-consuming task – irrespectively of the size of the programme – that requires capacity among implementers to monitor progress and to adopt intervention accordingly. Although rooted in extensive feasibility studies, the planning phase and the training of local partners took longer than expected partially due to low capacity on the ground, especially in the complex Liberia, Kenya and Mekong pilots. Delays in the design and planning translated also in further time pressure in the implementation phase. For example, in Mekong 15 villages were

added to the project after the baseline collection was fielded and the team decided not to collect any data for them, and left them out of the IE.

The experience with the RBI shows the importance of balancing time pressures against the benefits of good planning, of allowing enough time for implementation and adaptation, for a solid IE to be carried out and for learning. Setting up and testing the logistical arrangements, and feeding the findings back into the design of the intervention, take time, but it is invaluable to the success of the pilot. In some cases, existing work can feed into the process but, when prior evidence is scarce, it is also important to conduct focus group discussions, or field-test the intervention. For example, in Peru, an experiment focusing on training related to the RBI intervention was launched just before the programme and served to inform and improve the programme (Karlan and Valdivia 2011).

4.4. *Lesson 4: focus on the recruitment stage*

The recruitment stage is critical to the success of interventions that are inherently small scale because a further reduction in sample size lowers the probability of finding statistically significant impacts. Interventions focusing on business development skills appear particularly prone to recruitment problems (McKenzie and Woodruff 2012). The RBI were not exceptions, and they all faced difficulties in recruitment, except Liberia, where the group was pre-selected and there was no control group. There were problems in recruiting enough firms (Egypt, Kenya and Peru), finding enough villages (Mekong) and randomising among them.[9] The level of the randomisation and the final sample size is presented in Table 5.

A number of factors may lead to difficulties in recruitment including the time that targeted potential beneficiaries take to learn about the intervention, mobility constraints that may limit the accessibility of the interventions and an underestimation of the potential pay-offs of the intervention by potential beneficiaries. It is essential for teams to allocate enough resources and time to adequately address these challenges.

However, it is also important to acknowledge that participating in the treatment has a cost including foregone labour earnings, and potential beneficiaries may rationally decide that the expected pay-off is not enough to cover the participation cost. Teams should then see recruitment difficulties as a flag for the relevance of the programme and/or its accessibility for targeted beneficiaries and consider alternative designs.

4.5. *Lesson 5: select indicators early and make sure they are measurable and meaningful*

Selecting appropriate performance indicators is a challenging but critical step for the evaluation stage because it influences not only the measurement but also the multiple channels through which impacts transpire. It is essential to select indicators early during pilot design and ensure they are measurable and meaningful, keeping in mind that measurable changes can only be expected where there is scope for improvement, and if sufficient time is given for the different processes to take place.

Measuring impact on business outcomes is a complex task due to measurement errors in firm profits, revenues and sales (de Mel, McKenzie, and Woodruff 2009) and because the intervention can affect the reporting of business outcomes (McKenzie and Woodruff 2012). Pilots on female economic empowerment face additional challenges in measuring impacts on agency, both as a final goal and as instrument for better economic opportunities.

Table 5. Randomisation, clusters and sample size.

Country	IE method	Cluster	Stratification/matching Variables	Number of participants	Cluster and individual take-up rate
Egypt	Randomisation of matched pairs.	Firm	Economic sector, size, gender composition of work force and gender policies.	16 firms, 1040 employees.	80 per cent firms; 59 per cent employees.
Kenya	Randomisation of matched pairs.	Village	Village size, distance to a main road, beadwork and IGA experience.	23 producer groups, ranging from 7 to 70 members.	Initially: 50 per cent groups; after additional recruiting: 58 per cent groups.
Liberia	First difference (comparing results before and after).	n/a		246 women of Ganta's producer group.	n/a
Mekong	Randomisation of matched pairs.	Village	Village population, district location, percentage of households in BHP, type of BHPs, type of traders.	43 villages (or BHPs), 986 BHP households.	85 per cent villages due to nonrandom assignment to treatment and control.
Peru	Randomised control trial.	Individual	Districts/neighbourhoods.	1983 female microentrepreneurs.	51 per cent women.

Note: BHP = bamboo handicraft product; IGA = income generation activities.
Source: Authors'compilation based on Golla and Jones-Demen (2011); Golla and Saggers (2011); Golla and Selim (2011); Golla (2011); Knowles (2011) and Valdivia (2011a, 2011b).

It was particularly challenging for the RBI due to the limited evidence then available on measuring empowerment and identifying levers to enhance it (Narayan 2005). The pilot developed several innovative approaches relying on questions that are innately subjective.[10] But not all indicators proved meaningful or measurable, and in some cases inconsistent answers between men and women, or over time, were recorded.

4.6. *Lesson 6: supplement quantitative data with quality qualitative information*

Good-quality data is another essential element of a successful pilot, and using mixed methods provides a broader understanding of the results and associated processes, especially with limited budgets.

Reliable quantitative data require a large sample to minimise measurement error, and collection of quantitative information on some indicators – consumption, empowerment – requires high capacity of local agencies and enumerators. It also requires time – especially if new questions/indicators are introduced – because questionnaires need to be tested and some variables need time to be collected or to show changes. Qualitative data can usefully complement quantitative information and reduce the time and resources needed to meet the quality standard. Qualitative data collected at a preliminary stage and thorough testing of the survey instruments can help provide evaluators with better understanding of the baseline levels of key indicators and therefore provide a basis for comparison.

All RBI pilots, except Liberia, collected both quantitative and qualitative data. The IE of the Mekong RBI was particularly successful in using mixed methods in the evaluation. The quantitative questionnaire included eight questions (with up to eight answers) on decision-making about bamboo handicraft products (BHPs) and several questions on division of labour within the household. The qualitative work was instrumental in interpreting the findings on gender roles. In Laos, it indicated that there was a large increase in the percentage of BHP-related decisions made by females, and women felt in control of the money generated by the sales of BHP, which they used mostly to buy food in the market.

4.7. *Lesson 7: measure impacts at level at which change occurs*

It is essential to collect data at the level at which changes occur even though this may require a larger sample size. If the project works primarily to create change at enterprise (group/village) level, for example, impact should be measured and data collected at that level.

The RBI pilots collected data at several levels and in most cases measured the impact at the level the change was expected to occur: individual, household, producer group or firm. In Egypt, where the impact was expected at firm level, data were collected with a firm questionnaire – administrated to the human resource department of each firm – and an employee questionnaire – to human resource department and workers. However, in Kenya and Liberia, due to problems during the implementation stage, the evaluations focused on individual-level impacts, despite group-level implementation of the interventions. This discrepancy may partially explain the statistical weakness of the results in these countries.

4.8. *Lesson 8: monitor progress along the way and continue to do so*

Another main lesson is the importance of choosing indicators that change at different times and planning for several rounds of final data collection. There is no simple rule on the optimal time frame for IE. The optimal lag in collecting end-line data depends on the theory of change – that is on the type of changes expected and on the performance

indicators chosen. Enhancing female economic empowerment is a complex process that requires time and capacity among implementers to monitor progress regularly and to adapt the intervention accordingly. Indicators of final impacts should be complemented with immediate and intermediate impact indicators that can be monitored within shorter time frames and tend to be easier to measure. These indicators are also useful to gain information on the transmission channels and on emerging barriers that may prevent final impacts from materialising. Moreover, in this way, the long-term trajectory of the impact can be identified (Woolcock 2009).

Most RBI pilots collected end-line data within 6 months from the end of the intervention. In retrospect, this time frame was too short to be able to observe the full impacts of the interventions, let alone the trajectory of impact, especially given their ambitious objectives and the performance indicators selected. In Peru, for example, initial end-line data showed impacts on sales only for women who benefited from both the general training and the technical assistance, while the indicators on intermediate impacts – bookkeeping, enrolment in business/financial associations – showed positive impacts for all beneficiaries. A second round of data collected a year later mostly confirmed the results on intermediate outcomes but also showed a slightly higher increase in sales (Valdivia 2011b).

4.9. Lesson 9: if you are aiming for generalizable results, be consistent

Programmes that aim for generalizable results require consistency in design, implementation and IE methodologies across interventions. Adapting interventions to reflect country context and needs threatens the generalisation of findings. Using different IE methodologies weakens the robustness of results. Thus, teams need to weigh the trade-offs between reflecting country contexts and obtaining lessons that hold across interventions in line with the objective of the overarching programme.

The RBI varied considerably in design, choice of beneficiaries, implementation and IE methodology. The cross-pilot variation proved to be too high to allow for generalizable conclusions to be drawn. Given that this was a major objective of the programme, more weight should have been given to having consistency at all stages of the interventions.

4.10. Lesson 10: remember just-in-time monitoring

Finally, the role of just-in-time monitoring in avoiding pilot failures should not be underestimated. Monitoring is part of the due diligence to ensure that resources are spent adequately. It also allows for early detection of problems and, if combined with effective feedback mechanisms, for real-time fine-tuning of the programme, or possible early exit. This is particularly important for pilots because their innovative features may require more adaptation during implementation than more established interventions. Finally, monitoring can contribute to reconstructing the story behind success/failure and understanding the underlying transmission mechanisms (Bamberger, Rao, and Woolcock 2010; Rao and Woolcock 2003).

All RBI included a monitoring system, and a detailed common protocol was produced. In practice, this was rarely followed however, and the monitoring efforts varied across pilots in design, feedback mechanisms and depth of information collection. For example, the Kenya monitoring plan included several monthly collected indicators – including increase in individual/group sales and business turnover. In Peru, more detailed records on progress were provided through in-depth interviews with selected beneficiaries. However, generally regular monitoring fell significantly short of what initially envisaged, real-time

Box 1. A checklist to avoid pilot failures.

- □ Budget aligned with expectations and resources enough to incorporate learning component in addition to financing pilot intervention
- □ Number of subjects in sample large enough to provide results that are statistically significant at population level if IE is part of the design
- □ Timeline of the intervention does not compromise good planning and learning allows for adaptation if needed
- □ Potential recruitment issues analysed and tackled, and adequate time allocated to inform potential beneficiaries
- □ Enough time allocated between the intervention and the end-line data collection to allow processes to take place.
- □ Provision for collection of quality quantitative and qualitative data, including multiple rounds.
- □ Data collection envisaged at the level of expected impact, including multiple levels if needed
- □ Wide range of intermediate and final impact indicators which are measurable and meaningful
- □ Individual components of umbrella programme designed and implemented to obtain comparable and generalisable lessons.
- □ Monitoring system designed to provide real-time feedback on progress and system in place to feed findings into intervention design and implementation.

feedback was limited and never fully incorporated in programme adaptation. This may have been one of the major factors behind the failure of the RBI to generate the results they were set up to achieve.

5. Towards a checklist to avoid pilot failures

Pilot interventions are a growing feature of innovative approaches to development. However, the experience of the RBI shows that pilot programmes need careful planning, implementation and follow-up in order to provide useful results. Even small shortcomings and deviations from a few simple 'golden rules' may result in large departures from the initial objectives of the programme and may lead to pilot failure. Box 1 maps the 10 lessons learnt from this experience – and those not learnt – to corresponding items in a simple checklist designed to help practitioners avoid further costly missed opportunities.

Acknowledgements

This article draws on the experiences, results and lessons learned from the implementation of the RBI. The RBI were among the first projects designed as part of the Gender Action Plan (GAP). The RBI in the Arab Republic of Egypt, Kenya, Liberia, Mekong and Peru were funded by a Development Grant Facility grant to the UN Women and the ICRW. UN Women (then called UNIFEM) was responsible for the design and implementation of these RBI, while ICRW was responsible for the IEs; the Bank had an overall supervisory role. Mayra Buvinic, Lucia Fort, Andrew Morrison and Waafas Ofosu-Amaah from the World Bank conceived the initiative while developing the GAP. Involved in the implementation of

the Development Grant were Waafas Ofosu-Amaah (World Bank), Hiska Reyes (World Bank), Joanne Sandler (UNIFEM), Letty Chiwara (UNIFEM), Caroline Horekens (UNIFEM), Anne Golla (ICRW) and Anju Malhortra (ICRW). Several persons from the three organisations were involved in each of the pilots: Lorena Barba (UNIFEM; Peru), Helene Carlsson Rex (World Bank; Mekong), Carmela Chung (UNIFEM; Peru), Maria Elizabeth Dasso (World Bank; Peru), Izeduwa Derex-Briggs (UNIFEM; Liberia), Elisa Fernández (UNIFEM; Egypt, Kenya, Liberia, Mekong, Peru), Lucia Fort (World Bank; Peru), Anne Golla (ICRW; Egypt, Kenya, Liberia, Mekong, Peru), Caroline Horekens (UNIFEM; Egypt, Kenya, Liberia, Mekong, Peru), Zebib Kavuma (UNIFEM; Kenya), James C. Knowles (World Bank consultant; Mekong), Andrew Morrison (World Bank; Egypt), Maya Morsy (UNIFEM; Egypt), Sahar Nasr (World Bank; Egypt), Greg Ngungi (Consultant UNIFEM; Kenya), Waafas Ofosu-Amaah (World Bank; Kenya, Liberia), Ruth Okoth (UNIFEM; Kenya), Ryratana Rangsitpol (UNIFEM; Mekong), Hiska Reyes (World Bank; Egypt, Kenya, Liberia, Mekong, Peru), Meredith Saggers (ICRW; Kenya, Mekong), Asa Torkelsson (World Bank; Kenya) and Martin Valdivia (GRADE, World Bank consultant; Peru). This article benefited from comments by Jesko Hentschel and Mattias Lundberg (peer reviewers), Stefan Agesborg, Elena Bardasi, Jeni Klugman, Andrew Morrison, Waafas Ofosu-Amaah and Hiska Reyes.

Notes

1. Conversely, a pilot showing programme success does not guarantee that the programme can be successfully replicated in a different setting or scaled up. These issues are beyond the scope of this article, however.
2. A glossary of the few technical terms used here can be found in Gertler et al. (2011).
3. Only five programmes have been implemented and fully evaluated: Arab Republic of Egypt, Kenya, Liberia, Mekong Valley and Peru. Three programmes were added at a later stage and are still to be evaluated: Ghana, Nicaragua and Tanzania.
4. These findings are discussed in more detail in Johansson, Paci, and Posadas (2014).
5. Randomisation was affected by leakages and contamination between the treatment and the control groups given the proximity of villages.
6. For more details, see Knowles (2011) Table 1.
7. The expenses for data collection and M&E were: US$110,624 in Egypt, US$102,732 in Kenya, US$ 89,995 in Liberia, US$302,084 in Mekong and US$165,674 in Peru. Figure 1 in UN Women (2011) summarises the implementation costs per country: US$226,285 in Egypt, US$305,000 in Kenya, US$315,000 in Liberia, US$640,000 in Mekong and US$290,411 in Peru. The cost of the evaluation analysts was not itemised.
8. In the end, the number of clusters was maintained, but the number of observations per cluster remained below the minimum.
9. In Peru, the implementing team decided to expand geographically to the South Cone of Lima after failure to adequately reach eligible microentrepreneurs due to poor response in the North Cone.
10. The Demographic and Health Surveys are one exception, but they mainly focus on questions about decision-making.

References

Bamberger, M. 2009. "Strengthening the Evaluation of Programme Effectiveness through Reconstructing Baseline Data." *Journal of Development Effectiveness* 1 (1): 37–59. doi:10.1080/19439340902727610.

Bamberger, M., V. Rao, and M. Woolcock. 2010. *Using Mixed Methods in Monitoring and Evaluation*. World Bank Policy Research Working Paper No. 5245. Washington, DC: World Bank Group.

de Mel, S., D. McKenzie, and C. Woodruff. 2009. "Are Women More Credit Constrained? Experimental Evidence on Gender and Microenterprise Returns." *American Economic Journal: Applied Economics* 1 (3): 1–32.

Clark, M., R. Sartorius, and M. Bamberger. 2004. *Monitoring and Evaluation: Some Tools, Methods, and Approaches*. Evaluation Capacity Development Working Paper, Report No. 24614. Washington, DC: World Bank.

Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." Chap. 61 In *Handbook of Development Economics*, Vol. 4, edited by T. Paul Schultz and J. Strauss, 3895–3962. North Holland: Elsevier. doi:10.1016/S1573-4471(07)04061-2.

Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings, and C. M. J. Vermeersch. 2011. *Impact Evaluation in Practice*. Washington, DC: World Bank.

Golla, A., and M. Saggers. 2011. Kenya Results-Based Initiatives Program: Strengthening Export Competitiveness of Women Bead Workers. International Center for Research on Women, mimeo.

Golla, A. M. 2011. Impact Evaluation for the Small-Scale Project: What Do Managers Need to Know? Lessons Learned from the Results-Based Initiatives (RBI) Program. International Center for Research on Women, mimeo.

Golla, A. M., and A. Jones-Demen. 2011. Results-Based Initiatives Program Liberia: Value-Added Cassava Enterprise for the Ganta Concern Women's Group in Liberia. International Center for Research on Women, mimeo.

Golla, A. M., and M. Selim. 2011. Egypt Results-Based Initiatives: Promoting Gender Equality and Productivity in the Private Firms in Egypt; The Gender Equity Model Egypt. Impact Evaluation Report, International Center for Research on Women, mimeo.

Johansson, S., P. Paci, and J. Posadas. 2014. *Lessons Learned and Not Yet Learned from a Multicountry Initiative on Women's Economic Empowerment*. World Bank Study No.83276. Washington DC: World Bank.

Karlan, D., and M. Valdivia. 2011. "Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions." *Review of Economics and Statistics* 93 (2): 510–527. doi:10.1162/REST_a_00074.

Khandker, S. R., G. B. Koolwal, and H. Samad. 2010. *Handbook on Quantitative Methods of Program Evaluation*. Washington, DC: World Bank.

Knowles, J. 2011. *Final Report: Impact Evaluation of the Mekong Results-Based Intervention, mimeo*. Washington, DC: World Bank.

McKenzie, D. 2010. "Impact Assessments in Finance and Private Sector Development: What Have We Learned and What Should We Learn?" *The World Bank Research Observer* 25 (2): 209–233. doi:10.1093/wbro/lkp011.

McKenzie, D., and C. Woodruff. 2012. *What Are We Learning from Business Training and Entrepreneurship Evaluations Around the Developing World?* World Bank Policy Research Working Paper No. 6202. Washington, DC: World Bank. doi:10.1596/1813-9450-6202.

Narayan, D. 2005. *Measuring Empowerment: Cross-Disciplinary Perspectives*. Washington, DC: World Bank.

Rao, V., and M. Woolcock. 2003. "Integrating Qualitative and Quantitative Approaches in Program Evaluation." In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, edited by F. J. Bourguignon and L. P. da Silva, 165–190. New York: Oxford University Press.

UN Women. 2011. Results-Based Initiatives Program Evaluation: Report on findings. Authored by Woller, Gary, UN Women, mimeo.

Valdivia, M. 2011a. Training or Technical Assistance for Female Entrepreneurship? Evidence from a Field Experiment in Peru. World Bank, mimeo.

Valdivia, M. 2011b. Training or Technical Assistance for Female Entrepreneurship? Evidence from the Second Follow-Up for the Peruvian Field Experiment. World Bank, mimeo.

Woolcock, M. 2009. "Toward a Plurality of Methods in Project Evaluation: A Contextualised Approach to Understanding Impact Trajectories and Efficacy." *Journal of Development Effectiveness* 1: 1–14. doi:10.1080/19439340902727719.

World Bank. 2006. *Conducting Quality IEs under Budget, Time, and Data Constraints*. Washington, DC: World Bank.

World Bank. 2011. *Gender Equality and Development: World Development Report 2012*. Washington, DC: World Bank.

Appendix

Table A1. Mapping RBI features to best practice principles of impact evaluation.

	Justifiable	Theory of change	IE questions	Performance indicators
Arab Republic of Egypt	Can affect a large number of workers if scaled up; similar interventions implemented in other countries.	Gender equitable practices in the workplace: • Give women more job satisfaction → increase productivity. • Result in better match between tasks and skills → increase productivity.	• Gender equitable practices reduce gender discrimination in recruitment, provide on-the-job training, career development and prevents sexual harassment. • Prospect of gender equity certification provides incentives for gender equitable practices in workplace.	• Firm level: Proportion of female employees, proportion of women in newly hired, proportion of women in management. • Employee level: Job satisfaction measures, perception of (i) openness of work environment, (ii) willingness to promote and (iii) fairness of supervisor, compensation and training opportunities.
Kenya	Knowledge gap on the effect of training on women's business skills.	• Training contributes to developing women's business skills → increase productivity. • There are economies of scale in markets → women in groups have more access to inputs' and outputs' markets.	• Training and mentoring increase women's business skills. • Business skills increase productivity. • Group participation increases market awareness. • Larger control over income → increase in women's bargaining power within the HH.	• Group level: Business practices of the group (sales, trading and so forth) and revenues. • Individual level: Income-generating activities of women and group participation.
Liberia	Evidence of effect of training on female farmers' skills scarce.	• Training and mentoring contribute to develop farming skills → higher productivity. • Access to start-up capital → higher productivity.	• Training increases women's farming skills. • Access to capital increases women's productivity.	

Mekong	Knowledge gap on the effect of training on women's business skills.	• Training and mentoring contribute to develop business skills → increase productivity. • Increase productivity in bamboo production → more diversified household income.	• Business skills increase productivity. • Larger control over income → women's bargaining power within the HH. • Larger control over income → higher investment in children's education and health.	• Participation in bamboo production. • Sales. • Children's educational attainment. • Food consumption. • Decision-making by HH members.
Peru	Knowledge gap on women's business skills.	• Training and mentoring contribute to developing women's business skills → increase productivity.	• Business skills → higher productivity. • Larger control over income → increases women's bargaining power within the HH.	• Business practices. • Sales. • Time spent in enterprise. • Access to credit.

	M&E	IE method	Targeting	Sample size	Time frame	Data collected
Arab Republic of Egypt	Yes	• Randomisation of matched pairs of firms. • Matching on a few characteristics and with very few firms → control group could differ in many unobservables.	• Clear definition of eligibility rules: Medium and large exporting firms in greater Cairo. • Recruitment problems: Very few firms replied to initial recruitment → concerns about selection and external validity.	• No power calculations and very small sample size, particularly for cluster corrections.	• Short time to observe changes in managers' behaviour and workers' cultural attitudes.	• Quantitative workers' survey. • Quantitative firm survey.
Kenya	Yes	• Randomisation of matched pairs of villages.		• No power calculations.		
Liberia	Yes	• Before and after.	• A specific group of producers.	• No power calculations.		

(*Continued*)

Table A1. (*Continued*)

	M&E	IE method	Targeting	Sample size	Time frame	Data collected
Mekong	Yes	• Randomisation of matched pairs of villages.	• Three provinces identified through feasibility study. • Villages with producer groups and bamboo traders; but not all control villages had producer groups.	• Below the recommended size by power calculations for cluster corrections, by village or producer groups.		• Household quantitative data; FGDs with producer groups, IDIs with traders and village leaders (or key informants).
Peru	Yes	• Randomised control trial of female microentrepreneurs.		• Within the recommended sample size by power calculations.	• BDS and TA span 3-month period each, and about 6 months after endline data were collected.	• Interviews with female microentrepreneurs. • FGD with beneficiaries. • IDIs with municipal officials.

Note: BDS = business development skills; FGD = focus group discussion; HH = household; IDI = in-depth interviews; TA = technical assistance.
Source: Authors' compilation.