Review of M&E Practices in Mercy Corps' MSD Programs

Mercy Corps has been implementing MSD programs for over a decade, yet currently there is little guidance provided by the agency on how to approach M&E for MSD programs. Through our ongoing support for MSD programs, the TSU are aware that our programs have been adopting a variety of monitoring and evaluation approaches, with mixed success.

In July and August 2018, the Markets TSU team conducted a review of M&E practices for past and current MSD programs, led by an intern (Dana Juha). The objectives of the review were to better understand the approaches and tools that our MSD programs have been using, and to identify best practices and common failings. It is intended that findings from this study will provide the basis for internal guidance for field teams, in terms of recommended approaches and resources for MSD programs.

Methodology

There were two main steps in the process of undertaking this study. Firstly, a review of global best M&E practices for MSD was undertaken, in particular focusing on the Donor Committee for Enterprise Development (DCED) standards but also drawing on other resources such as the Adopt-Adapt-Expand-Respond (AAER) framework and the BEAM Exchange. This was then used to develop a list of key practices against which Mercy Corps' programming could be compared. Practices were divided into the following categories:

A. Use of Results Chains
B. Measuring the baseline
C. Measuring systemic change in the market system
D. Measuring impact on beneficiaries
E. General practices

Secondly, the study reviewed the M&E practices of thirteen MSD programs, of which six were completed and seven of which were on-going. The review of each program entailed interviewing key personnel and analyzing program documents and measurement plans. The following table shows the final list of reviewed programs (Figure 1).

Figure 1. Use of Results Chains

Completed Programs	On-going Programs
Alliances ALCP - Georgia 2008-17 (2017-21 ongoing)	IMAGINE - DRC
EC SWITCH - Indonesia	ARC - Jordan

DALIP - Kenya	M-RED - Nepal
MVMW-Myanmar	Google Labs - Jordan
E4A - Timor-Leste	Li-Way - Ethiopia
Rain - Uganda	Light up - Liberia
PRIME - Ethiopia	

Findings

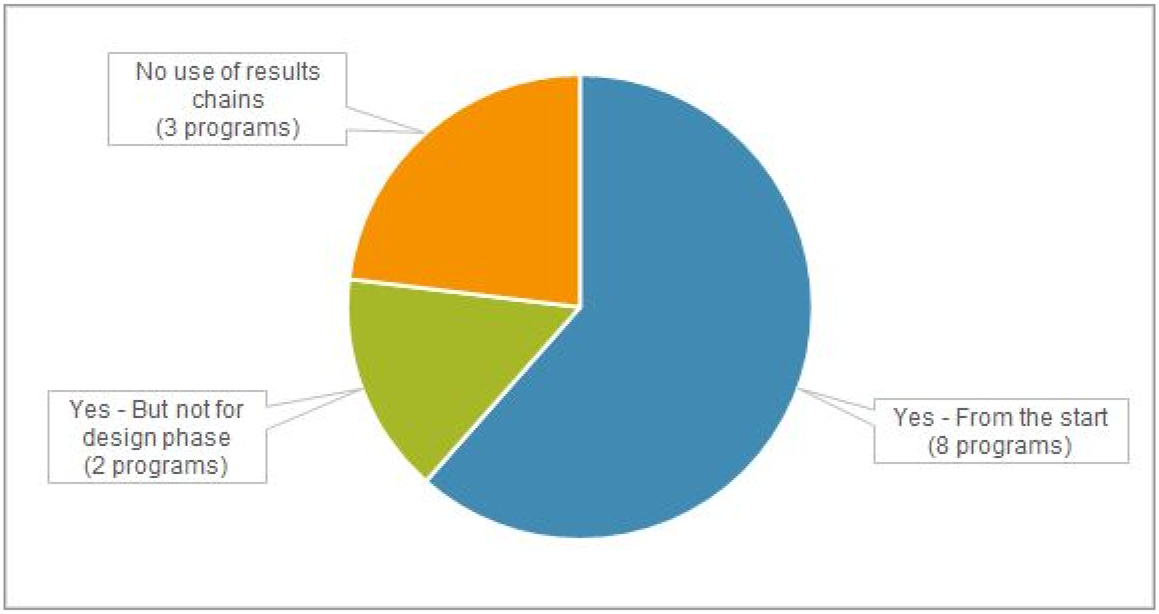
This section presents the main findings from the review, with respect to each of the five categories of M&E practices. The report focuses on aggregated findings across programs and highlights the nature of the challenges faced. When relevant, particular programs are used to describe positive examples to ensure lessons can be learned from their documentation and tools.

A. Use of Result Chains

1. **A surprising number of Mercy Corps' MSD programs have not used Results Chains at all, or did not use them for design and strategy.**

   Five out of the thirteen interviewed programs (37.5% - see Fig.2) either did not use Results Chains at all (three programs), or did not use them in developing their program strategy (two programs - one of which introduced Results Chains at the mid-term, and the other used Results Chains only to design their final evaluation). This is not a shockingly high number, but given the importance of Results Chains in all aspects of MSD implementation and results measurement, it shows improvement is still needed. In particular because the sample of programs the survey covered are considered Mercy Corps' 'best' MSD examples.

Figure 2. Use of Results Chains



2. **In many cases, use of Results Chains has been driven by either the donor, an external partner, or by the TSU.**

   Of the ten programs that used results chains, four programs adopted them because it was a requirement of the donor or pushed by a third technical partner, and three programs adopted results chains due to technical guidance from the TSU. The other three programs can be said to have introduced them independently. Mercy Corps cannot always rely on external agencies to push the development of Results Chains and needs to ensure all of our MSD programs use Results Chains as a core foundational tool.

3. **Most of the surveyed MSD programs use program-level Results Chains (rather than more detailed activity-level Results Chains)**

   Out of the ten interviewed programs that used Results Chains, seven used the equivalent of program-level Results Chains. These programs spoke positively about how Results Chains helped their program and teams. The level of detail was sufficient to push the team to develop a clear vision of their intended system change and program logic. It was also sufficient to link to their results measurement framework (with the exception of activity-level indicators).

4. **Three of the surveyed MSD programs used activity-level Results Chains.**

   The LI-WAY program, the Alliances program and the RAIN program all used activity-level Results Chains. All three programs agreed that the process for going into that level of detail is time consuming and requires staff training (for example Li-WAY's approach was driven by the donor and had the support of Springfield consultants). Interestingly, both LI-WAY and Alliances chose to include only the key activities in the Results Chain, rather than a comprehensive breakdown of every step. This seems like a good compromise to the problem of overly burdensome activity-level Results Chains.

5. **Most of our MSD programs have used Results Chains to develop their performance measurement plan (PMP), however this is not necessarily a formalised process.**

   Seven of the eight programs that used Results Chains in their program design, also linked their Result Chain to their results measurement framework (or PMP in Mercy Corps terminology). Some of the programs formalised this process, explicitly including indicators for each box in the Results Chain, while others simply referred to the results chain to help inform indicators. Making the link explicit is preferable, as it forces us to measure system change at every step.

   One exception to this was the ARC program. In this case the donor requested for the PMP to be developed in the early stage of the inception phase, and the team did not have the flexibility to significantly change the indicators after it had been approved. In ARC's case, market assessments and result chain activities were conducted after the development of the PMP, largely because the donor required the PMP at an early stage.

6. **While many of the MSD programs include a focus on gender, in practice gender is very rarely included in the development of Results Chains.**

   Li-Way and Alliances are the only two programs that incorporated gender and cross cutting groups in Result Chain design. Li- WAY is already a gender-based program and activities revolve around gender considerations. Alliances gave a good example of distinguishing gender sensitive activities in their Result Chain, where activity boxes related to gender are given a different color.

7. **Challenge: Staff training and building a culture of adaptive management was a key challenge for a number of MSD programs.**

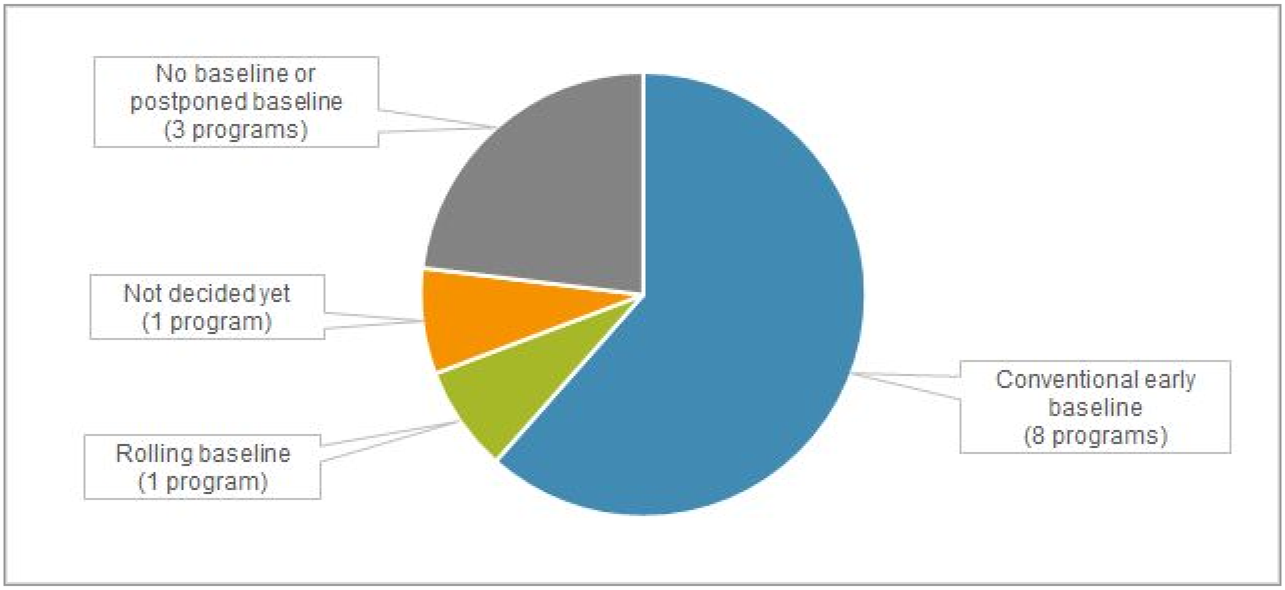
   Managers identified that communicating to program staff about how a Results Chains should be adapted over time versus a static document was a challenge. Given the time and effort teams put in to the initial drafts of these documents, there is some reluctance to adjust over time. However, most programs found the tool useful to understand the logic of the interventions and explain the project to new staff members who joined after the design of activities.

B. Measuring the Baseline

8. **How to approach baseline surveys is a major source of uncertainty and confusion for Mercy Corps' MSD programs.**

   A common tendency is for our MSD programs to conduct a baseline right at the beginning of the program, using some variation on systematic random sampling as one would for a conventional 'direct-delivery' program. Eight out of the thirteen surveyed programs took this 'Conventional Early Baseline' approach (Fig. 3) and this caused significant problems for results measurement, as will be described below. In most cases this was an automatic decision made in the absence of any specific guidance about alternative approaches, though for some programs (e.g. ARC) there was also pressure from the donor to conduct the baseline very early. One of the programs (Li-WAY) was still at a very early stage of implementation at the time of this review and had not yet decided how to approach their baseline.

Figure 3. Baseline approach



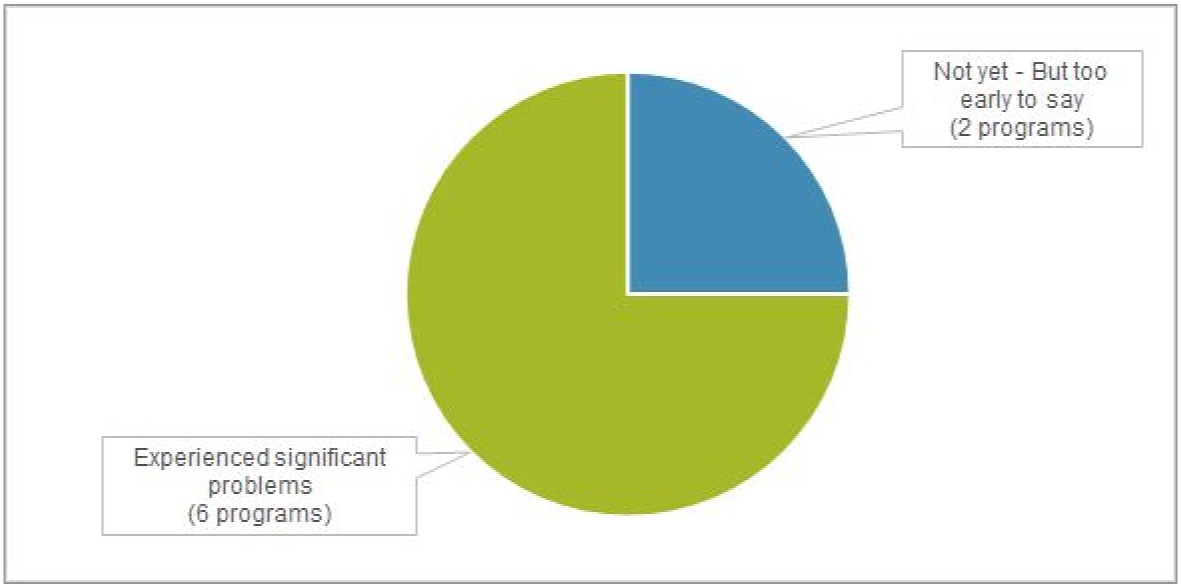
9. **Programs that conducted a 'Conventional Early Baseline' experienced numerous problems with the data it produced.**

Of the eight surveyed programs that used a ‘Conventional Early Baseline’ approach, six experienced significant problems. The other two have not yet conducted their endline and it is too early to say whether they also have problems with the data that their baseline produced (see figure 4).

The major challenge that programs encountered was that the participants in their baseline survey were not reflective of the final beneficiaries ultimately reached through the program. This meant that a longitudinal panel study was not possible (comparing specific beneficiaries over time), and also that comparing a population sample at endline with the baseline population sample was inherently unreliable as it meant they were comparing different population groups (essentially ‘comparing apples with oranges’). There were three main reasons this problem arose:

- Firstly, because the survey was undertaken very early, before the programs were designed in detail, it was not possible to predict which population groups would be impacted and in what way. This resulted in missing key target populations (e.g. one program was not able to predict when conducting the baseline that it would be working with small-scale dairy producers and thus has no baseline for these actors).
- Secondly, baseline figures became an issue when a program significantly expanded its geographical reach due to market actors replicating new practices, a common occurrence in an MSD program. This meant that no baseline existed for the new geographies and these new populations (e.g. one program expanded significantly when it shifted from being a conventional ‘direct-delivery’ program to using a facilitation approach).
- Thirdly, when programs impacted an unexpected population group in addition to the primary target group (e.g. one program expected to impact SMEs, but ended up having a larger impact on low-wage employees working in them, and had no baseline data for this group), baseline results also proved challenging.

Figure 4. Problems with ‘Conventional Early Baselines’



Another common problem related to ‘Conventional Early Baselines’ is that the questions asked in the baseline do not adequately capture the changes that we ultimately want to see in the program. Again, this is the inevitable result of the baseline being conducted too early, as it is impossible to predict key questions related to market system change before the program strategy and results chains have been developed. This is not so much a problem for very high-level impact level indicators (such as household overall income or food security), but it is a big problem when asking questions related to market uptake and behaviour change (e.g. adoption of new technologies or production practice) or more specific impact indicators (e.g. income from a particular economic activity - such as dairy sales).

10. **The ‘Conventional Early Baseline’ problems led our MSD programs to take numerous steps to remedy its shortcomings**

In order to overcome the failings of the ‘Conventional Early Baseline’ approach, the surveyed programs resorted to a number of measures:

- At least one program re-did the baseline at the mid-term evaluation. The final evaluation data was ultimately compared with the mid-term data rather than the baseline data, so in this instance the baseline assessment was redundant. One other current program is expected to follow a similar approach.
- Several programs ended up using retrospective reporting by beneficiaries in their endline survey to capture impact, rather than comparing against the baseline. Again, the baseline assessment was redundant as a measurement tool.
- One program introduced an additional process of regular annual surveys to capture impact on beneficiaries over time, which was separate to their large baseline and endline surveys that measured impact on nutrition and food security.

11. **Two programs (Alliances and DALIP) provide examples of how MSD programs can take a more strategic approach to baseline assessments.**

The Alliances program has been through many iterations since it started in 2008. The first program was supported by Springfield Centre, and under their guidance it was decided not to conduct a baseline at all. Instead, at the end of each phase, Alliances conducts a comprehensive Impact Assessment in which beneficiaries and non-beneficiaries are surveyed and report on their various practices and performance. This includes retrospective reporting for the two previous years, enabling Alliances to measure improvements. This is triangulated against previous impact assessments and official statistics.

The DALIP program took a slightly different approach. They postponed their baseline until they had begun implementation, and then conducted four different baselines for the four different intervention areas they were working in. They identified baseline participants using information from the private sector businesses they were already partnering with by this point. This approach therefore did not in theory rely on retrospective reporting by households. However, apparently no endline survey was ever conducted to compare against this baseline, and instead impact was captured using qualitative questionnaires. As this process was led by the donor (Kenya Markets Trust - led by ASI), it wasn’t possible to get more information on this from the Mercy Corps team.

It’s also worth noting that one other on-going program (Google Labs) has also chosen not to undertake a baseline assessment. This is in part because of the nature of the sector they are working in (youth employment at scale).

12. **Most MSD programs chose to design baseline tools internally and outsource baseline data collection to an external consultant.**

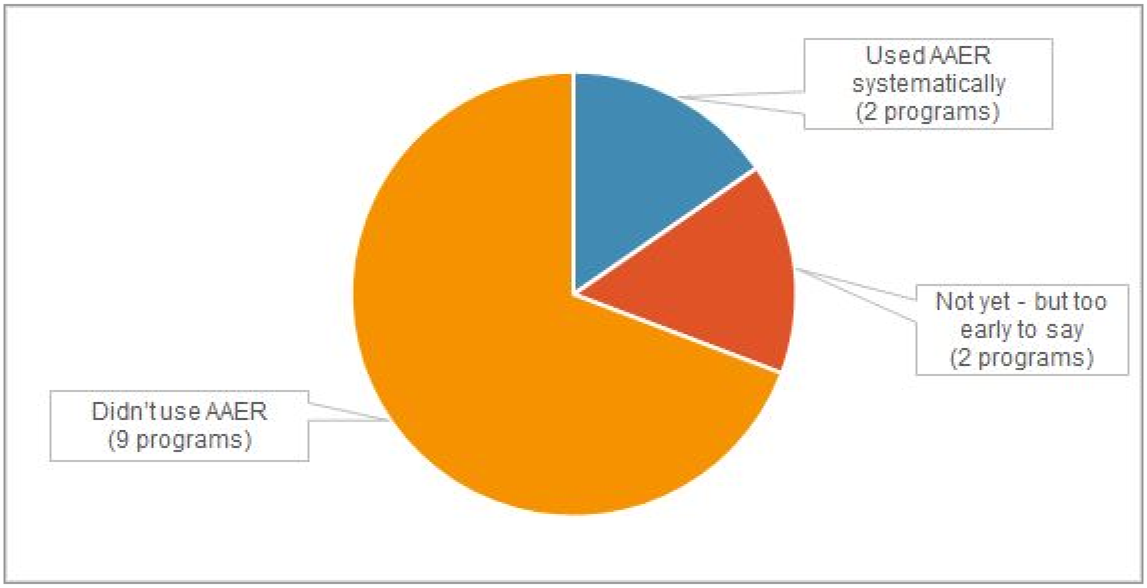
    Seven out of the eleven programs that conducted a baseline, outsourced to an external consultant. MVMW's first baseline was a rolling baseline, therefore it was conducted internally. For the second baseline which targeted a different state, a conventional random sample baseline was conducted due to the time and effort needed for the first rolling baseline.

C. Measuring Change in the Market System

13. **Very few of our MSD programs have used the AAER framework to systematically analyse and measure systems change.**

    Nine of the thirteen surveyed programs did not use the 'Adopt, Adapt, Expand and Respond' (AAER) framework for analysing system change, and most of these had never heard of it (Fig. 5 below). Two programs are still in the early phases of implementation and it is too early to say whether they are using AAER systematically, though both said they are intending to use it.

Figure 5. Use of the AAER framework



14. **Most of our MSD programs are doing a pretty good job of analysing and reporting on the 'Adopt' component of the AAER framework for systems change.**

    All of the surveyed programs documented the number of partnering businesses that had adopted new behaviours or business practices ('Adopt'). In a few cases, there was a tendency to focus on the number of business supported, rather than rigorously analysing whether this support had led to a sustainable change. But most programs were using measures such as 'repeated business behaviours' and 'additional business investment' as indicators to show that the new practices were sustainable.

15. **While some of our programs make an attempt to capture more complex system change (e.g. Adapt, Expand and Respond), most do not do this in a systematic and comprehensive way, making it hard to verify 'attribution'.**

Quite a few of the surveyed programs reported some evidence of partnering businesses using their own initiative and resources to introduce additional new products, services or practices ('Adapt') or non-partnering businesses crowding-in ('Expand'). But with the exception of two programs, described below, there was no comprehensive system in place to capture these types of system change systematically. In many cases these findings were based on somewhat anecdotal evidence, without any rigorous process in place to dig into whether these additional system changes were stimulated by Mercy Corps' MSD program. For the Adapt, Expand and Respond components of system change, we are therefore not able to claim 'attribution' with confidence.

There is also a need to embed system change formally in our Results Measurement systems. For example, Results Chains can include specific steps related to businesses innovating independently 'Adapt' and crowding-in ('Expand'). Performance Measurement Plans (PMPs) also need to do a better job of including indicators related to these multi-dimensions of system change. For example the MRED program included the following indicators for measuring change in non-partners, forcing them to measure this systematically:

- *"Indicator 3.3.1: # of government agencies and private sector actors that replicated best practice examples promoted by Mercy Corps."*
- *"Indicator 3.3.2: # of communities outside MRED target areas that have replicated or expanded piloted mitigation projects with their own resources or have plans to do so at end of project (MRED 1 MI 2.2)"*

16. **Two programs (Alliances and DALIP) incorporated more comprehensive approaches to measuring system change.**

The Alliances program provides an excellent example of how a team can put in place systems to strategically measure systemic change. The program had a staff member specialized for measuring systemic change. The process was described as iterative and required having a strong network of partners. The team collects both quantitative and qualitative data regularly, debates these findings rigorously every month, and documents findings on a systemic change log every three months. The Systemic Change Log is a comprehensive tool used to report the program's systemic change attribution.

The DALIP program used the AAER framework which was driven by the donor. AAER measurements had to be part of the program's quarterly reports. "Adopt" and "Adapt" used to be measured and documented regularly by the field staff, looking into partners' sales, level of investment and introduction of new products. "Expand" was captured through asking partners about their competitors and changes in the market. The team had a steering committee that worked closely with the government and was used to measure the "Respond" piece.

17. <u>Challenge:</u> **Willingness of private sector actors to share information is a key challenge for Mercy Corps programs that attempt to measure systemic change.**

Eight programs reported facing challenges to systematically collect information from either their direct partners or non-partners. Private sector actors are not always keen to share sensitive information related to their business.

D. Measuring Impact on Beneficiaries

18. Our MSD programs generally have good systems in place to measure the number of beneficiaries reached (breadth of impact).

All of the programs used information provided by private sector partners to collect data on the number of beneficiaries the program had impacted. Most asked private sector partners to provide information (such as sales records or training records) to program staff on a monthly basis. In the case of the Alliance program, this monthly data collected from private sector businesses extends to information about the businesses' clients (beneficiaries), such as sales.

Five of the six completed programs primarily used this data collected from private sector partners to estimate the number of beneficiaries they had reached. Only Alliances used a different approach; they used their endline impact assessment to estimate how many beneficiaries had been impacted, by asking beneficiaries whether they had accessed one or more program-facilitated service or good.

Figure 6 below shows the number of beneficiaries reached by each of the surveyed MSD programs that had been completed (data was not available for on-going programs).

Figure 6. Measurement of numbers of beneficiaries reached

Name of program	Reported number of beneficiaries impacted *How this data was collected*
E4A	More than 10,000 households purchased improved energy products as a result of the program *Measured using sales data from partnering distributors and retailers or energy products, including retail businesses crowding-in.*
SWITCH	773 businesses producing tofu and tempeh (SMEs) purchased new energy efficient production technology. *Measured using sales data from manufacturers and distributors of the efficient production technologies.*
DALIP	5,763 households accessed services facilitated by the program, including index-based insurance and agro-vet services. *Measured using sales data collected from private sector partners*
MVMW	More than 25,000 farming households purchased and used new inputs or services and 9,471 households increased their value of production by more than 25%. *Measured using sales information and training information from partnering input retailers and service providers.*
RAIN	More than 61,000 farmers accessed goods or services facilitated by the program. *Measured using information collected from agri-processing companies, financial service providers and many other private sector actors. They used an auto-generated unique identifier approach to calculate overlap, with all*

	beneficiaries characterised by name, village, age and gender. This data was also triangulated against endline survey data.
Alliances (Phase 2014-17)	64,000 farming households accessed at least one program-facilitated good or service. *Measured using a large-sample endline Impact Assessment, and triangulated against data collected from private sector partners.*

19. **Most of the surveyed programs also did a good job of measuring the extent to which beneficiaries were impacted (depth of impact), despite the challenges caused by using a 'Conventional Early Baseline' approach.**

The problems with the 'Conventional Early Baseline' approach were described in Lesson 9 above, and some of the measures programs took to overcome these problems was described in Lesson 10.

Despite these challenges, all six of the completed programs were able to report on the way that key characteristics of beneficiaries had changed in the lifetime of the program (Fig. 7).

Figure 7. Measurement of impact on beneficiaries

Name of program	**Reported impact on beneficiaries** ***How this data was collected***
E4A	Households purchasing energy products reduced energy-related expenditure by 70% on average, resulting in a savings of $78 per household per month. *Measured using an endline survey with beneficiary households reporting retrospectively (through recall). This was triangulated against the baseline survey and laboratory tests of technologies (e.g. for cookstoves).*
SWITCH	● 292 tofu and tempeh factories (38% of beneficiaries) increased their profits, by an average of 25%. *Measured using an endline survey of producers, who reported retrospectively on business profits.* ● On average, tofu and tempeh factories reduced their 'per production unit' energy consumption by 86% and 74% respectively. *Measured using field tests of technologies, which were triangulated against MSE survey data.* ● More than 2,000 workers in tofu and tempeh factories increased their wages by an average of 22%. *Measured using an endline survey of producers, who reported on businesses costs including wages paid.*
DALIP	DALIP didn't measure impact, the program was handed back to Kenya Market Trust (KMT) at the end of the program.
MVMW	9,471 farming households increased their value of production by more than 25%, and 8,483 farming households increased their income by more than 50%. *Measured by comparing the sample of beneficiaries surveyed in the endline assessment with the sample of beneficiaries surveyed at the baseline.*
RAIN	More than 30,000 households increased their income (52.4% of beneficiaries). *Measured using beneficiary recall compared with the previous harvest season.*

	This was backed up by analysis based on 'calculated income' of beneficiary farmers (in locations where baseline data existed).
Alliances (Phase 2014-17)	40,000 households generated a tangible positive income change, generating net additional attributable income of $4,734,048 USD. *Measured using a large-sample endline Impact Assessment with beneficiaries reporting retrospectively, and findings triangulated against impact assessments from previous phases and against official government statistics.*

Despite these successes, it's important to note that the problems with the initial baseline not only created extra work for our MSD programs (for example through having to conduct a second baseline), but also undermined our ability to report on impact. For example, programs that had to conduct second baselines later in the program, or compare their endline with the mid-term instead, ended up only capturing impact for part of the program, rather than from the beginning, which likely meant under-reporting. In other cases, programs had to rely on beneficiary recall to capture change, which can work quite reliably in some contexts (such as the Alliances program in Georgia where beneficiary recall was triangulated against additional programmatic impact assessment findings and official government statistics), but can be very unreliable in other contexts.

20. **A weakness of almost all of our surveyed MSD programs has been a failure to quantitatively measure attribution when analysing beneficiary impact.**

The Alliances program was the only program that comprehensively measured for attribution when reporting on beneficiary impact. They did this by measuring the counterfactual, and comparing data for user-groups (beneficiaries that used at least one program-facilitated service or good) with non-user groups (non-beneficiaries that hadn't used any of these services). This was collected in the endline impact assessments, using beneficiary recall from two previous seasons.

The PRIME program also attempted to measure attribution quantitatively, but faced an additional challenge that all households in the chosen geography had been exposed, to one degree or another, to program interventions. Instead, they categorised households by 'high intensity' and 'low intensity' treatment, and compared outcomes of these households at baseline, midline and endline.

Some of our other MSD programs used alternative methods to try to show how market changes had impacted beneficiaries. For example, the RAIN program used an Outcome Harvesting approach to first identify positive outcomes, and then trace these back to their drivers. For other programs, however, this was qualitative only.

The importance of being able to measure our attribution quantitatively is highlighted by the example of the Alliances program. In one region of Alliances implementation, the 2016 Impact Assessment showed that income for beneficiaries (livestock and dairy producers) fell by 18% between 2014 and 2016, the result of a border closure for six months that was crucial for livestock trade, and also due to a major devaluation of the local currency. Without measuring the counterfactual, and quantifying attribution, this would have looked like a failure of the program. But in fact, Alliances were able to show that for *non-beneficiary* households, average income fell by 45% over the same period. By measuring attribution in this way, Alliances was able to prove to a high degree of statistical certainty, that beneficiary households were more resilient to shocks in the economy, and that the program led to a large increase in net additional attributable income (over $1,670,000 USD in 2016).

E. General M&E Practices

21. **All Mercy Corps' MSD programs that reported the use of Result Chains manage their measurement plan interconnectedly.**

Eight programs have their measurement plan managed by M&E staff and available for all program members on a shared drive.

22. **All Mercy Corps' MSD programs consider M&E activities interconnected with program activities and ensure the involvement of all team members.**

Mainly program or field staff are responsible for data collection and following up with partners. Program teams meet on a regular basis (monthly/quarterly) to review Result Chains and triangulate data collected in the field.

23. Programs that are taking a mixed programmatic approach, faced challenges in integrating MSD M&E tools with non-MSD components.

MRED and IMAGINE have non-MSD interventions in their program. IMAGINE is facing challenges in measuring attribution of their MSD activities against the direct implementation approach of their other interventions. MRED is facing challenges explaining to the donor the constant change that is happening to their MSD interventions.

24. **It is likely that our MSD programs are under-budgeting for M&E.**

For most programs, M&E was highly integrated into program activities. As a result, it was difficult for the surveyed programs to give an exact percentage of their budget that was allocated to M&E activities. Based on our discussions, and anecdotal feedback from programs, it seems that our MSD programs tend to under-budget for M&E capacities and practices. This is backed up by broader analysis of M&E spending by Mercy Corps programs, which shows our spending on M&E is far lower than our agency target of 5%.