

Learning from Failure

Executive Summary

Driven by a wish to learn more from what goes wrong in our programming, and to examine where changes to the broader organization and system can improve our programming and impact globally, in 2019 CARE undertook its first evaluations-based failure meta-analysis. This analysis draws learning and evidence from 114 evaluations of CARE's work from 2015-2018 to understand the patterns and trends in what goes wrong. This helps us take a data-driven approach to strategic investments and action plans to live out CARE's commitment to high program quality and continuous improvement across the board.

The review draws from project specific data, but deliberately anonymizes the data and focuses on overarching trends to remove blame for any specific project team or set of individuals. This exercise is designed to help us learn more about how we can change our processes and patterns of support and engagement around weak areas to improve our work.

Methods: The analysis applies qualitative research methods to a subset of CARE's project evaluations where the evaluator cited areas that we could improve or obstacles that may have slowed or reduced project impact. This subset represents roughly 15% of CARE's evaluations from the relevant timeframe (2015-2018). It is important to note that within this subset, most of the projects were successful—achieving a majority of their targets and demonstrating significant impact. Only 8 projects in the sample (i.e. 7%) failed to meet most of their goals. Examining failures across a range of projects, even successful ones, gives a better sense of which challenges are common, systemic, and difficult to overcome. The team used MAXQDA, a qualitative analysis software, to examine reports for failures according to a pre-determined set of 61 codes grouped into 10 overarching themes (see Annex 1).

The way forward: We are already working within CARE to address the challenges that this report highlights, and to focus on teams and individuals who can take action that will help unlock these larger trends. Over the next months we are having conversations with a variety of stakeholder groups in the organization to further examine the nuances and implications of findings within their specific areas of responsibility so we can create and monitor action plans for progress. Further, we are working to develop ways of measuring trends and diagnosing failures more rapidly, so we are not only drawing from learning at the end of a project cycle. We hope to run this analysis again in mid-2020 to identify progress and new areas of focus.

Introduction

This meta-evaluation complements an ongoing effort at CARE to focus more on learning from failure—including our [Failing Forward podcast series](#) and country, project, and team-specific conversations about learning from what went wrong. Those efforts have so far focused on case studies that look at an individual project's challenges and learn from the details of that specific experience. That rich detail is critical because it allows us to think more critically about adaptive management and addressing failures as they are happening. It also allows people to take action within their own projects to avoid common mistakes.

To further strengthen our learning from failure, this meta-evaluation provides a more holistic look at trends across projects, regions, and sectors to see where we may have structural challenges that contribute to the challenges project teams face. By taking this more systemic view, we hope to be able to prevent or reduce future failures by making broader improvements to our organization, our support structures, and our processes. This system-wide approach moves from a focus on individual actions and challenges that seem isolated to one project or context to a bigger consideration of trends that affect many people and projects across CARE. While individual actions and reflections are critical to learning from failure and improving our work, learning is maximized when backed by broader support from the whole organization, and shifts in the enabling environment teams work in.

This meta-analysis draws lessons from what evaluators reported as challenges, gaps, and lessons learned across 114 project evaluations from 2015-2018. Failures were identified as things that went wrong and varied greatly in severity, from delayed implementation and Monitoring, Evaluation, Accountability, and Learning (MEAL) planning, to projects lacking a Gender Based Violence (GBV) plan. We hope that this report can galvanize constructive conversations around failure, allowing CARE to better serve vulnerable populations around the world.

This report is based on 114 of CARE's project evaluations from 113 projects spread across regions, sectors, and member partners. The sample consisted of reports written between 2015-2018 where the evaluation discussed challenges, obstacles, inefficiencies, or lessons learned. While there were identifiable failures in 113 of the 114 analyzed evaluations, this does not suggest that the projects themselves were failures. Many met or exceeded the majority of their targets and goals, contributing to significant and sustainable impacts for people across the world. Of the 114 evaluations, only 8 were for projects that significantly failed to achieve their goals. The rest represent projects that were able to meet most of their targets, but where the evaluation highlighted challenges or obstacles they faced along the way that may have slowed or reduced progress.

We selected primarily final evaluations across CARE's body of work that were written between 2015-2018, although in a few cases we used mid-term reports where there was no final evaluation available when those evaluations showed particularly rich insight into what goes wrong. Evaluations were coded and analyzed using MAXQDA and then exported to Excel for further analysis, visualization, and dissemination. We looked at 61 specific codes that fell into 10 overarching-categories: HR, Budget, Partnership, Implementation, MEAL, Scale, Gender, Design, and Other. On average, evaluations had failures in 4.40 of these 10 categories. For this report, we are focusing on the MEAL and gender categories, two critical areas to CARE's mission and vision.

Given the qualitative nature of this research, findings are not generalizable across CARE's body of work; rather we hope to highlight failure trends in a systematic way to complement the work being done to improve project design, implementation, and evaluation. CARE closes approximately 300 projects a year, so we expect to have about 900 projects that ended in the time frame this report covers. From that set of possible projects, we selected for projects where there was a final or midterm evaluation, and specifically selected the sub-set of evaluations where evaluators raised challenges and offered suggestions for how to improve the work. All statistics in this report should be viewed in that context: of the evaluations where we saw challenges, a certain percentage of those evaluations contributed to each coding category. This is an important distinction from claiming that a certain percentage of CARE projects were failures or failed in a particular area.



CARE team practicing a learning from failure exercise

This data will be used for iterative learning through transparency and learning around key failures in HR, Budget, Partnership, Implementation, MEAL, Scale, Gender, and Design by region, and to propose specific recommendations and develop timelines to address points of improvement. CARE is working with relevant stakeholders across the organization to generate action plans that will help us address these failures moving forward, and continue to improve our efficiency and impact.

Methodology

Given the descriptive and varying nature of CARE's evaluations, we felt that a qualitative analysis best captures project failure to promote learning. The team established an initial set of codes and an Excel sheet was created to consolidate evidence of those failures. As the project grew, the team turned to qualitative analysis software to more effectively capture work. Originally, the team used RQDA, however, after numerous technical difficulties and problems sharing work, in January 2019, the team transitioned to MAXQDA. At this point, the codebook was revised and adapted based on learning from previous coding. For the full codebook, see Annex 1.

Coding: To design the coding system for the meta-evaluation, the team began by listing a common set of implementation failures that they had seen in their own project experience and from having initially read the evaluations with an eye to summarize impact data. The codes also draw inspiration from the priorities outlined in CARE's Program Strategy and the components of OECD's effectiveness and efficiency criteria. These codes were split into larger categories (HR, MEAL, etc.) to correspond with key CARE priority areas and stakeholder groups in the organization that can take action to correct failures. Once the initial list of codes was developed, the team then tested it against 10 random evaluations from the sample set to gauge its robustness and validity. At that time, code adjustments were made to better reflect the trends present in evaluations. Additionally, the team circulated the revised code book to a set of CARE's MEAL experts for further refinement. There were minor improvements made at this stage; however, since the full coding exercise was a completely new experience, there was limited visibility into which codes would be most useful. Having done the coding for the full evaluation set, we now have much better insight into what codes we would use for future rounds of the project.

The codes relating to gender were particularly problematic, and did not align as closely with the evaluations as

the other areas. For example, nearly half of the codes identified under gender fell into the category "miscellaneous"--reflecting gaps in the initial analysis and forecasting. After a review of those original codes, the qualitative data behind them, and in consultation with CARE's Gender Justice technical team, we proposed a new set of codes for gender that better capture the content of the document. Those new codes are reflected in this document.

Sampling: All CARE evaluations are publicly available on CARE's Electronic Evaluation Library, per CARE's Evaluation and Accountability Policy. Drawing from this repository of information, the team selected 114 evaluations based on the following criteria:

- The evaluations were completed between 2015 and 2018, and represented the most recent report available on any given project (this was primarily final evaluations)
- The evaluations discussed challenges, lessons learned, or failures as part of their content.
- A stratified sample reflected a variety of regions and sector areas across the organization to show the most holistic picture possible, and to allow for more specific analysis with stakeholder groups like regional management units and technical teams that can influence system-wide change to improve program quality over time
- The evaluations were either in English or French--the two languages the coding team could read
- The file size, format, and availability could be imported into MAXQDA for qualitative analysis

Review process: The 114 evaluations were divided among the three team members for coding. Segments of text in the evaluation reports were coded anywhere an evaluation indicated a failure, such as preventable and unforeseen challenges and limitations. Several codes could, and often were, applied to the same segment. The team also utilized the memos function in MAXQDA to share questions, comments, and reflections through the coding process. Work was then merged every couple of weeks into a common database. Once MAXQDA was introduced, the coding process took roughly 215 hours over 3 months, split between the three coders. However, one coder had previously read many of the reports, decreasing her raw coding time. Merging, organizing, and cleaning the data took an additional 50-60 hours over the course of the project. We intended to review all codes as a team to decrease inter-coder subjectivity, however, due to time constraints, we were not able to do this, and reviewed only categories where we anticipated discrepancies, such as repercussions.

What is Failure?

For the purpose of this report, a failure does not represent that projects failed to meet their goals. Rather, failure is identified as foreseeable or preventable challenges as identified by the evaluators. Therefore, identification of failure is based on the evaluator's perspective on both what challenges existed in a project and what warranted attention in a final report. Thus, failures are likely to be skewed to issues that were never solved and failures related to project impact rather than the systems and processes to which the evaluator might not have access. For example, evaluators rarely have access to detailed project budgets. Often failures appear in the "lessons learned" or "challenges" section in a final report, however, most were interwoven throughout the report, making it difficult to gain these insights without a detailed reading of the document. We selected a range of projects in terms of how successfully they met their goals to move away from the idea that only bad projects fail.

Analysis and Reporting: At the end of coding, before analysis, documents were organized into sets based on region and sector. The regions included were: West Africa (WARMU), Middle East and North Africa (MENA), Latin American and the Caribbean (LAC), ECSA (East Africa), Asia (ARMU), and Global/Regional, which constituted mutually exclusive categories. The sectors were: Resilience, Governance, Education, Water, Sexual and Reproductive Health Rights (SRHR), Women’s Economic Empowerment (WEE), Food and Nutrition Security (FNS), and Humanitarian, with several evaluations falling under multiple sectors.

Once coding was finished, all the information was exported directly from MAXQDA and organized in Excel. We added columns to identify the projects’ sectors and imported a page on code co-occurrences (i.e., how many times any two codes were both applied to the same segment). Then, key indicators were extracted to highlight global trends and quantitative information. We examined overall figures on failure trends as well as co-occurrences of failures to gain insights into how certain failure categories can act as triggers. While the purpose of this report is to highlight our qualitative findings, these quantitative figures have been included to depict failure trends across CARE’s work.

In this report, all project-identifying information has been removed to focus on broader trends; the aim is not to shame or blame projects, but rather contribute to a culture of learning. This includes translating all quotes into English as necessary for the purposes of this report.

Adapting our codes

Within the 114 evaluations analyzed, the number of coded segments per evaluation ranged from 0 to 137, with 2,777 initially coded segments. After coding was finished, we added an additional code called “**Design**” as a final overarching category to all failures that could have been addressed during the project design phase. The addition of the design category identified 1,031 additional coded failure segments, resulting in 3,808 total coded segments. Often, evaluations would repeat the same challenges and barriers in multiple locations, therefore, a single issue could be coded several times throughout the evaluation. Furthermore, segments could and often were coded for several different overlapping failures (e.g., both data quality and data analysis), thus the 3,808 codes does not represent 3,808 separate failures. Therefore, it is not appropriate to add the different categories and feeder codes together, as there is significant overlap in the codes.

The original **gender** category had seven feeder codes. What made this category particularly interesting was the high occurrence of the miscellaneous gender feeder code, which had the highest number of coded segments (59 out of 195 gender-related codes) and was present in the second highest number of evaluations (26 out of 57). While this likely indicated inadequate code development on our part, it also points to the complex nature of gender related failures.

With this insight that the gender codes were least able to reflect what evaluators were highlighting, we worked with CARE’s Gender Justice team to develop a new list of codes which better reflect the nuances of the field. This new set of codes draws from CARE’s Gender Equality and Women’s Empowerment Framework, which looks at three main pillars that can support a women to realize her full human rights. These pillars are:

- Agency: a woman’s skills, hopes, and abilities that she needs to succeed;
- Structure: the environment that surrounds and conditions her choices, including laws, policies, and social norms;
- Relations: the power relations through which she negotiates her path, especially related to her connections with family, friends, and other members of her community.

The codes further draw from our gender design standards to articulate what we hope to see from projects, and where projects often go wrong. The updated codebook (Annex 1) reflects these new standards.

Limitations

Through the research, there were various limitations affecting efficiency and effectiveness, however, these were vital to learning and provide important lessons for future failure analyses. Technology proved a persistent problem from the outset. Initially, we recorded failures by copying segments into a master excel. This soon became difficult to navigate, thus, we sought out qualitative analysis software. First, we used RQDA, a free qualitative analysis software powered by R Studio. However, information in RQDA could only be analyzed in raw text format. This meant coding was time consuming and often key pieces of information were missing, such as graphs, diagrams, and headings. MAXQDA proved a much more effective tool for the purposes of this project, however it is more expensive, and the number of uses is limited to the number of licenses purchased. We recommend that further failure analyses expand access to allow for further efficiency.

These technological limitations greatly reduced the time available for the project analysis. Unfortunately, the project was limited to only four months which included: codebook development, coding, and analysis. The project would have benefited from more time dedicated to codebook development to synthesize some repetitive codes and further clarify definitions. After coding, the amount of time left to analyze was also too short. As mentioned above, it was originally intended for each code to be reviewed by multiple people to ensure consistency throughout, but this was not possible. Failure analysis would benefit from an iterative coding process for the first few months to strengthen coding definitions and ensure intercoder reliability.

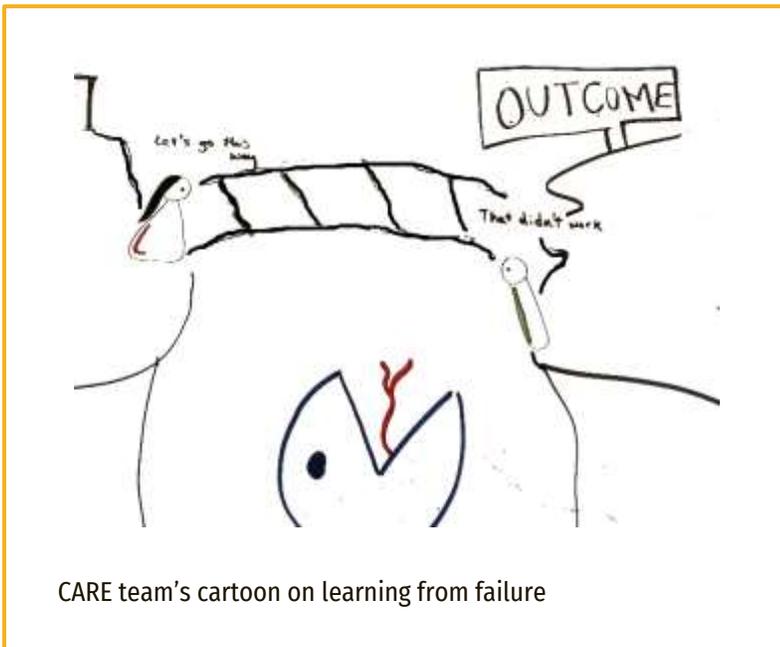
Another important limitation in the study is that we did not include any evaluations written in Spanish, Portuguese, or Arabic. Evaluations in these languages make up 7% of CARE's global evaluation database, but no reviewer on this project was able to read in those languages, forcing us to exclude them. This necessarily also reduces the sample size from the LAC and MENA regions.

The final significant limitation was related to the evaluations themselves. Evaluations varied in focus and content. Failures across similar projects could have been represented differently depending on the style of evaluation and the evaluators' opinion. Some evaluations were written in an overly optimistic or positive manner. For instance, one project evaluation mentioned that the project faced "challenges", however, without

any specific mention or details on what those challenges were, no segments were coded. Conversely, other evaluations were pessimistic despite quantitative results highlighting significant positive development. For example, some evaluations emphasized the fact that implementation areas produced worse results than the control areas, with limited acknowledgment to that fact that other interventions were introduced in those control areas. Therefore, evaluators presented this as an implementation failure, rather than a MEAL failure.

Reporting bias was evident in comparing the qualitative and quantitative reports from the same project. In this case, the qualitative report identified 65 codes segments for failure over 6 master buckets,

as well as design failures, whereas the quantitative report had only 5 coded segments, all with the MEAL category. These two evaluations were coded by the same person in the same day, eliminating the discrepancy



CARE team's cartoon on learning from failure

from intercoder variability, and limited the influence of intra-coder variability. Final evaluations are subject to the bias and focus of the evaluators, an important consideration when reviewing failure results.

Recommendations

Through this report, we hope to continue conversations within CARE on constructive ways to discuss failure to improve project implementation across regions and sectors. We want to build on the dedicated work of teams in Madagascar, Bangladesh, and South Africa which host Fail Fair and Failing with Honesty conversations, as well as the Water team that is experimenting with pre-mortem activities in their project design phase. It is our hope to further open spaces to discuss failure as a means of learning and adaptation to best serve vulnerable populations all around the world.

With this body of work, we can further dive into how projects are failing both within regions and sectors. We aim to produce learning briefs to further dissect and detail what failures look like within these specific categories, as understanding failure trends is the first step to addressing them. In particular, we want to dig deeper into our qualitative analysis of each of the overarching categories, to better understand how and why failures are happening for each of these and continue exploring the common co-occurrences for failures to better understand how one failure type can trigger additional failures.

We also recommend drawing from a broader group of reviewers with more diverse language skills so that we can draw from a broader and more representational set of examples for learning and action.

Lastly, we leave several recommendations and lessons learned for future coding endeavors. Primarily, we suggest having each coder analyze the same set of several evaluations before refining a finalized codebook for the project. This could help highlight areas of discrepancies in understanding codes to better ensure intercoder reliability. Establishing this clear foundation before starting requires a larger time contribution on the onset but will help mitigate confusion and discrepancies during the analysis stage. Furthermore, we simply had too many codes. Codes should be broader and more encompassing, especially within our Implementation category. Future teams should continue the practice of regularly uploading and important teamwork to mitigate the negative effects of technical difficulties.

Next Steps

The learning from this report provides many opportunities for learning and action to further improve our work. The focus of the exercise is not simply to highlight failures, but rather to give us a structured way to look at failure more systemically so that we can improve our work overall. Building from this learning, we can change some of our overarching processes and support systems to reduce common failure points for individual projects and teams. With that in mind, CARE is taking a number of steps to use this learning to prevent future failures.

Team-specific action planning: We are working with diverse teams around the organization to examine the sub-set of data most relevant to them and develop action plans for broader systemic change. For example, we hosted a conversation with the Food and Water Systems (FWS) team to present the data for FNS projects and discuss action steps for improving this work. The FWS team paired this data with a mini-Failure Fair

discussing specific case studies to get a better sense of specific challenges. From there, the FWS team is creating a set of next steps to address the issues.

We are also working with the staff members most closely involved in project design to identify the issues that often go wrong during the design phase and create tools and action plans to address those.

Improving our project design process will mitigate or prevent many of the most common failures this report identifies. Similarly, we will be working with the Gender Justice team and the CARE International Gender Network to understand what these findings imply for the tools and resources available to support projects in this area. Over the next 6 months, we are holding conversations like this around the organization to create tangible action plans for improvements in the areas that came up as challenges most often. We will be checking in on these action plans every 6 months to see if we have been able to sustain improvements.



Community member discussing what we could do better

Informing organizational investments: CARE is continually investing in improving our organizational performance, and the findings from this study reinforce the importance of these investments and will inform specific actions within these broader investments. Some examples of how this will work are:

- **Monitoring, Evaluation, Accountability, and Learning:** CARE is already investing in making impact measurement and learning a priority for the organization. Over the last five years we have built a global impact measurement system with organization-wide metrics. The system permits us to aggregate and compare results across different geographies and approaches to identify the most effective programming and scale up what works. This kind of global aggregation is atypical in the industry. We have committed more than \$5 million to invest in workforce skills development, technology upgrades, knowledge management and data-driven storytelling for advocacy. Part of this investment has included creating a CARE-wide MEAL working group for improving MEAL quality across the network, regional investments and working groups for MEAL, and supporting communities of practice. The learning from this meta-evaluation informs these investments and communities of practice to see which specific gaps we should be addressing that are impacting the broader system and help us use this investment most effectively to drive change.
- **Program Quality:** CARE has renewed our commitment to invest in program quality. The global teams that focus on program quality in country and sector settings work to provide tools and resources for projects to design projects more effectively and to respond to challenges as they arise. Some examples of this work include our efforts to measure and improve quality through the CARE Markers on Gender, Governance, and Resilience. The markers trigger annual sensemaking and action planning sessions to focus on improving program quality. This meta-evaluation provides critical insights into common obstacles that programs face, reflects on some of the underlying causes that may contribute to lower quality, and supports our global program teams in targeting their efforts to address cross-cutting risk areas for many projects.

Advocacy, Influence, and Scale: CARE is committed to sharing our learning and evidence with others so that everyone can improve their work and benefit from our experiences. This includes significant investments in advocacy, learning and sharing, and scale. While some of the challenges this report highlights are specific to CARE, some reflect broader challenges in the international development space. This structured reflection on systemic challenges—drawing from a dataset that cuts across donors, geographies, and sectors—provides useful opportunities to work with others to identify areas where we can all work together to improve impact in programming. The data in this report can be structured against the OECD quality standards and industry-wide definitions of efficiency to give many actors an entry point for understanding the implications of these findings. We would also be happy to share our methodology and approaches with any others willing to conduct a similar learning exercise in their own organizations.

Continued Improvement: We will be checking in on action plans bi-annually to ensure that we are improving our systems and support to teams based on what this report is telling us. Additionally, CARE intends to repeat this exercise every 2 years to hold ourselves accountable to improving our program quality—and to identify new failure points as they arise. Our hope is that within the next 2 years we will have successfully addressed many of these issues and seen a reduction in failures in our most common problem areas. One detail to note here is that many projects that will close in 2019-2020 have already been designed and begun implementation, so there is a limit to how much improvement we would expect to see in the Design category over the next 2 years. Given the typical program cycle of 3-5 years, those efforts would not show up in our dataset until 2022 or later.

ANNEX 1

Failures Codebook

HR

- **Hiring Delays:** It took too long to hire a staff member (or all staff members), and this resulted in problems with the project. Possibly staff positions were left vacant for significant periods of time.
- **Turnover:** staff regularly left a particular position and/or there were many positions where staff left resulting in a gap in skills, knowledge, or ability to get work done.
- **Staff Capacity:** staff did not have the critical skills needed to get the job done. Staff were missing some skills related to project technical design or implementation. There was no training for required skills or tools as part of the project (e.g., not enough rollout of financial management systems).
- **Staff Balance:** there was a lack of diversity amongst staff (not enough women, not enough local staff, not a good split between responsibilities and how many staff were assigned to particular technical areas or partners).
- **Upper leadership buy-in:** managers somewhere in the project or top-level country office leadership (or leaders of partners), were not sufficiently supportive of a critical piece of the project, so it didn't get done, or was not done well enough. This happens most often with gender, MEL, learning, and advocacy—but is certainly not limited to that.
- **No field staff buy-in:** people doing the last mile implementation or coordinating local implementers do not believe a particular issue is important or don't understand well enough to implement. Most commonly cited around gender equality issues, but certainly not limited to that.
- **Missing Key Staff:** the project did not hire for (or possibly fund or design for) necessary staff. Maybe we have far too few field staff to realistically carry out activities. Tend to see this a lot with gender and MEAL positions, but it's not restricted to those issues.
- **Poor performance management:** misalignment between what staff are held accountable for and what needs to get done. Anything that mentions re-setting staff goals, managerial issues, etc.
- **Miscellaneous HR**

Budget

- **Burn Rate:** Project did not spend its total budget in the time allocated or at the pacing planned for in the project. Most commonly labelled as spending delays, burn rate, and no-cost extension.
- **Overspend:** project spent more money than planned, or significantly overspent a specific line item for particular activities.
- **Transferring to partners:** there were difficulties getting the money to partners in time to implement activities. Possibly the financial systems didn't match up, or the reporting cycle was delayed and partners can't get new funds if they don't submit a report. [This should also show up as a partnership failure.]
- **Missing pieces:** we failed to budget for particular activities or personnel that would have been important for project success (often see this in gender, MEL, communications, and advocacy)

- **Can't reconcile:** The budget doesn't add up. We're not sure what happened with money, or there are significant gaps between reporting and planning. This can have happened at any time in the project cycle, even if it was corrected.
- **Budget Delay:** Delay of receipt of funding from donor that impacts the project.
- **Miscellaneous Budget**

Partnership

- **Coordination:** working with partners took a long time, was really challenging, or didn't work at all. There were misalignments between the visions and/or activities between different partners.
- **Accountability:** partners failed to deliver according to their agreements (this could be financial problems, not doing activities, not submitting reports, not meeting goals, etc.). CARE did not put mechanisms in place to see what partners were doing or did not correct problems in time. Alternatively, partners and participants had no way to hold CARE accountable to our commitments to them.
- **Delivery:** partners are not doing the activities the way they are supposed to, or are not reaching the people in the way the project set out. To count here, it must specifically say that the partners are having a problem, and not that it's an overarching project problem.
- **Capacity:** the partners are missing critical skills to deliver and/or did not get necessary training to do the work in the partnership agreements.
- **Missing key stakeholders:** we did not partner with everyone we should have, or there are key gaps in our partnerships and/or implementation that someone else should have filled. We see this a lot in reference to local governments as partners or private sector actors.
- **Miscellaneous Partnership**

Implementation

- **Targeting:** we worked with the wrong group of people (not the extreme poor or picked people who could not invest in what we were promoting). Alternatively, we set the wrong project goals and couldn't meet them.
- **Lack of Context Analysis:** Lack of pre-project research or misunderstanding or missing knowledge around key contextual factors, including lack of knowledge on other organizations/CARE implementing similar projects in the same area.
- **Design to implementation slip:** we had a good project design, but for some reason aren't carrying it out. Maybe we didn't budget for it. Maybe it took longer than we allocated. Maybe it was a lot more complicated than we thought and we didn't have the right staff to do it. Etc.
- **Integration:** there are a lot of components to the project and they are all (or some) happening separately in a way that compromises efficiency or effectiveness. Possibly we are putting a burden on the community by conducting too many activities at different times. Possibly we have too many partners repeating context analyses with slightly different tools or technical focus areas. Etc.
- **Quality:** we just aren't implementing the activities very well, or the activities we planned weren't very good. Maybe the training quality is terrible. Maybe the staff can only go to a community once every three months because we didn't assign enough staff, so there's not enough support. Maybe

we didn't fund transportation, so people can't go do the work. Maybe the activities never made a lot of sense to solve the problem.

- **Input Supplies:** We (or partners) committed to get a physical thing to the communities so they could implement part of the project and it didn't happen or didn't happen on time. Alternatively, we did get the inputs out, but they were of low quality. We see this a lot with seeds and agricultural tools, but might also happen with health supplies, training tools, promised grant money, construction materials, etc.
- **Missing key realities:** something about our project design went wrong because we missed a piece of the context. We forgot to factor in transport costs to a marketing scheme. We had people build stoves that were so heavy no one could move them. We supported an income generating activity that has little or no demand in the community where we promote it. We trained women over 40 in a technique that is mostly relevant for new mothers, and new mothers are on average 17 years old. Etc. This is distinct from the "context analysis" piece in MEAL because it is not necessarily calling out the lack of a study, but rather that we designed a program without taking realities into account.
- **Implementation Delay:** it took too long to get activities going. As a result, they didn't happen, didn't happen soon enough, or got done so poorly that there was little impact.
- **Missing key necessary activities:** we didn't factor in activities that we needed to do. Often comes up around gender, and especially engaging men and boys, but there are lots of contexts where an evaluation will say "should have done x, but it wasn't in the project design."
- **Miscellaneous Implementation**

MEAL

- **Need More Context analysis:** something wrong with baseline, gender analysis, market analysis, and/or stakeholder analysis. Possibly we didn't do some of the analysis at all. Possibly we focused on the wrong thing (e.g., took national data but only worked in one district that's an outlier, looked at where people can sell their goods, but not where they buy the supplies they need)
- **Partnership:** didn't get the right MEAL partners. Group/consultant/partner who was working on MEAL did not do it effectively.
- **Data Analysis:** we collected a lot of data but left it sitting in databases and didn't do anything with it. We can't articulate the key lessons out of the data or what the data said. Includes sampling (i.e., we didn't collect data on enough people or the right breakdown on participants vs non-participants to make claims on what did or didn't work).
- **Adaptive management:** we found problems during our MEAL work (like at mid-term), but didn't do anything to fix them. This is especially true if the final cites a continued problem carried over from mid-term.
- **MEAL planning:** we didn't plan for MEAL at all, or not enough. We picked the wrong indicators (or didn't set any indicators). We changed indicators partway through the project. Our baseline data collection tool has nothing to do with our endline one. We can't find baseline data to compare to endline, or we can't find the tools we used the last time. Etc.
- **Staff:** Something wrong about who we recruited for MEAL or we didn't have staff in place at all (this is a connection to the HR bucket and should reflect in both).
- **No follow through:** we had a decent MEAL plan, and then didn't do it. We recognized a problem with

MEAL but didn't correct it. We invested in tools and didn't use them.

- **Data quality:** data is inconsistent between tables and narrative. Numbers don't add up in columns. Someone has averaged averages. We can't tell where the data came from or verify any of it. The sample sizes are too small to be relevant (usually anything under 100).
- **No budget:** we did not budget for MEAL or didn't budget enough.
- **Need more risk analysis:** we did not plan for predictable risks in a situation. This comes up most often around climate change in agriculture projects, but might also refer to floods in a flood-prone area or political changes during election years, etc.
- **No crisis modifiers:** we knew that there were risks (like climate) but made no provision to change what we were doing in case those risks happened (like re-allocating money to emergency response in case of a flood.)
- **No learning processes:** there is not enough (or no) attention paid to how people learn from the work and/or data collection. This is usually cited specifically as a learning problem.
- **No KM systems:** data is not easy to find, is not being shared, and/or is not being produced in a way people can consume. There are no common platforms for staff to access and use information.
- **Miscellaneous MEAL**

Scale

- **No structural partnerships:** we didn't do anything to work with organizations that could scale—like the government or private sector—or we did it too late in the project to be effective.
- **Too expensive:** the project had an impact, but at a cost no other actor can replicate. This is usually called out specifically as the program being too costly. Look for cost-per-participant/beneficiary or cost-per-impact numbers.
- **Sustainability:** the project didn't have a sustainability plan and/or exit strategy. Alternatively, project tried to add one in the last six months of the program, which is too late.
- **Miscellaneous Scale**

Gender:

- **Engaging women and girls:** We didn't work with women and girls enough (at all, or enough). We structured activities in a way that made it hard for women and girls to participate
- **Engaging men and boys:** the project focused primarily or exclusively on women and did not focus on how men and boys are gatekeepers to power, so the men and boys held women back. This often comes up around mobility, resource use and GBV issues, but might be a lot of other things.
- **Participation:** made up of the categories of engaging women and girls and engaging men and boys. This is useful for examining overlap, and presenting higher-level analysis with fewer categories for specific audiences.
- **No gender analysis:** we didn't do a gender context analysis, or we missed a critical piece. We misunderstood something about the context as it related to gender (e.g., we helped women increase incomes and their husbands thought that meant they were prostitutes).

- **No gender advisor:** overlap from HR section above
- **No GBV plan:** the project didn't have a referral system in place to help women who suffered from violence in conjunction with project activities.
- **Agency:** Drawing from the gender equality framework, the project did not pay enough attention to women's individual skills, confidence, and capacity when designing and implementing activities.
- **Relations:** Drawing from the gender equality framework, the project did not pay sufficient attention to women's relationships with others, and how women are able to negotiate within these relationships to improve their lives.
- **Gender MEAL:** The project did not measure changes in power dynamics and equality between men and women. Alternatively, the indicators the project used were not appropriate to the change they were trying to produce, or were not sensitive enough to capture changes in the community.
- **Did not look at structure:** Drawing from the gender equality framework, the project focused on training women but didn't look at limitations beyond just the skills women have. Often see this in market connections, representation with government, mobility, etc.
- **Gender repercussions:** women experienced backlash or regressive changes in power dynamics as individuals and communities felt threatened by women's increasing power and attention to women.
- **Miscellaneous Gender**

Other Categories

- **Repercussions:** the project participants experienced negative effects as a result of participating. Women suffered more violence because there was not enough effort engaging men and boys. People lost money because they invested in a technology that didn't work. The project didn't do a cost analysis and so people had to pay to do what the project encouraged. General lack of ethics and consent, etc.
- **Technology:** technology failed in some way. It wasn't targeted toward the people we try to reach (or they couldn't access it). Connectivity was too bad to sustain using the tools. Men used women's cell phones to track them. Etc.