Bringing rigor to evaluations of large-scale programs to improve infant and young child feeding and nutrition: The evaluation designs for the Alive & Thrive initiative

Purnima Menon, Rahul Rawat, and Marie Ruel

Abstract

Background. The evidence base on the impact of large-scale infant and young child feeding (IYCF) and nutrition programs is limited, partly due to the challenges of rigorously evaluating complex programs including multiple interventions.

Objectives. To describe the process used to design Alive & Thrive’s impact evaluations in the three target countries and discuss the feasibility of developing contextually relevant designs adapted to the country-specific programmatic context.

Methods. The evaluation designs for Alive & Thrive needed to address several challenges. These included the selection of intervention components to evaluate rigorously; the identification of appropriate comparison groups in the context of rapidly scaling-up programs; the choice of impact indicators; addressing measurement challenges related to evaluating the impact of interventions targeted during the first 2 years of life on stunting; and developing methods and tools to assess implementation, utilization, and program impact pathways within evolving program portfolios.

Results. In Bangladesh and Vietnam, cluster-randomized probability designs are used for the impact evaluations; in Ethiopia, the impact evaluation uses an adequacy design. In all three countries, repeated cross-sectional surveys, 4 years apart, are used to measure impact, and appropriate age groups are sampled separately to capture change in the main impact indicators. In addition, theory-driven process evaluations are used to study factors that facilitate or prevent achievement of impact and scale.

Conclusions. We conclude that robust impact and process evaluations of complex, large-scale nutrition programs are feasible, but that early implementer–evaluator engagement and shared vision and motivation to establishing meaningful evaluations are essential.

Key words: Impact evaluation, infant and young child feeding, nutrition, process evaluation, program impact pathways

Introduction

Infant and young child feeding (IYCF) practices are a major contributor both to child survival [1] and to the prevention of child undernutrition [2]. Global estimates, however, show that poor IYCF practices are widespread in most regions of the world, particularly those with the highest burden of undernutrition [3]. Even though reviews [2, 4–7] suggest that interventions to improve IYCF practices are available and effective, the evidence for intervention effectiveness comes mostly from small-scale studies, with the exception of breastfeeding interventions. Large-scale initiatives to improve breastfeeding suggest that a combination of policy advocacy, mass media, and community-based interventions led to changes in breastfeeding practices in Bolivia and Madagascar [8] and in other countries [9]. However, there are few examples of successful large-scale interventions to improve complementary feeding.

One of the primary reasons for the scarcity of evidence from large-scale effectiveness trials to improve IYCF practices is the significant methodological challenge of rigorously evaluating large-scale programs. Challenges range from establishing proper comparison groups that allow attribution of impact to the programs, adjusting evaluation designs to implementation processes and timelines, assessing impact pathways and quality and fidelity of implementation, and allocating adequate financial resources for impact and process evaluations. These challenges, however, are not insurmountable, and there is increasing recognition that
an evidence base from well-conducted evaluations is urgently needed in order to inform programs and policies to improve IYCF practices and improve nutrition at scale. Indeed, recognition of the poor resourcing environment for global health and poverty has led to several calls for bringing more rigor to program evaluation from different quarters [10, 11].

In this context, Alive & Thrive, a 6-year initiative that aims to reduce stunting through improved IYCF practices at scale, was established as a learning grant in 2008 [12]. With the goal of implementing large-scale interventions using multiple platforms to improve IYCF practices in Bangladesh, Vietnam, and Ethiopia, the initiative was also charged with developing a measurement, learning, and evaluation (MLE) approach to generate a strong evidence base for future actions to improve IYCF practices. The MLE goal is one of four main objectives of Alive & Thrive, and is thus embedded into the initiative’s core business.

Alive & Thrive’s MLE goals, as articulated in the overall grant proposal, were to document the impact, cost, and cost-effectiveness of IYCF interventions implemented at large scale through Alive & Thrive’s activities; and to generate learning on how to achieve and replicate Alive & Thrive’s impact.

The first step in the MLE agenda was therefore to establish rigorous, yet feasible and relevant, impact and process evaluations that were tailored to the unique context of each target country’s program. This paper describes some of the main principles and steps used in designing the evaluation approaches for the three countries, focusing on how generic and country-specific constraints to designing evaluations of complex programs were addressed. The paper, which primarily focuses on the impact evaluation design for the initiative, concludes with reflections on how to balance the sometimes conflicting demands of researchers who seek to establish rigor in impact evaluation, and implementers who face the multiple challenges of real-life program implementation and rapid rollout and scale-up. A companion paper in this Supplement [13] provides a thorough methodological overview of the process evaluation approach used in Alive & Thrive. Neither of these papers reports any evaluation results; they are exclusively focused on the evaluation designs.

Methods

Program evaluations should ideally provide reliable and accurate answers to the questions: “Did the program lead to anticipated outcomes?” and “Why/how were these impacts achieved, or not achieved?” Evaluation designs should also be tailored to the level of inference required from the evaluation; designs can range from adequacy designs (documenting whether changes in outcomes are observed or whether overall goals are met without attempting to attribute changes to the program; e.g., a program goal to achieve final stunting rates of 15%, or a program goal to train a certain percentage of frontline health workers in IYCF), plausibility designs (aiming to establish that the impacts achieved are likely to be due to the program, above and beyond other external influences that might have caused the observed changes), or probability designs (establishing with a high level of statistical probability that impacts were attributable to the program) [12]. Although probability designs that allow adequate statistical inferences of links between the program’s inputs and the outcomes are often desirable, the evaluation design process needs pragmatic and realistic considerations of the implementation context, the goals and expected uses of evaluation findings to inform decision-making, and the cost and feasibility of different evaluation design options [13–16].

Given the paucity of large-scale effectiveness evidence on improving the spectrum of IYCF practices through community-based interventions, our objective in designing the overall evaluation for Alive & Thrive was to develop, where possible, probability designs, with linked theory-driven process evaluations. We anticipated that this approach would offer the best possible opportunities to generate answers to the primary questions the evaluation team was tasked with addressing:

- What impact did Alive & Thrive have on IYCF practices and stunting?
- How was this impact achieved, or not achieved?

Defining rigor in program evaluations

A rigorous design is defined as one that minimizes selection and information bias, controls for confounding factors, and attempts to rule out chance and secular trends as contributors to changes observed over time [12, 14]. An ideal evaluation design, therefore, allows one to attribute changes in outcomes to the intervention that is put in place. This is best done by comparing changes over time between groups or individuals who are exposed to the interventions and groups or individuals who are not [15]. However, a major challenge to attributing impact to specific interventions or programs, even when such evaluation designs are established, is the issue of the identification of an appropriate comparison group [15]. The right comparison group is one that has the same chance as the intervention group to receive the interventions and, thus, is one that is not affected by selection bias. This is largely ensured by randomly allocating a set of individuals or geographic clusters to either the intervention or the comparison group. An important consideration here is that the random allocation itself be done correctly.

Additional challenges to ensuring overall evaluation rigor are that for programs or interventions that
focus on systems-level interventions with long causal pathways between programmatic inputs and impact indicators, there is a need for a theory-driven evaluation design that carefully considers and unpacks the intervention inputs, intermediary outcomes, and impacts and thus documents the pathways to impact [11, 16–19]. Last, but not least, the evaluation of nutrition interventions raises further challenges related to the biology of the nutrition outcomes being studied or the behavioral aspects of communications-focused interventions.

In designing the evaluations of Alive & Thrive, the overall aim was to consider all of these elements that must come together and to bring best practice in program evaluation [20, 31] to large and complex program models aimed at reducing stunting through improved IYCF practices in three target countries.

Overall approach to the evaluation design process

We drew on lessons learned from impact evaluations of large-scale social safety net programs and smaller-scale nutrition programs conducted by the International Food Policy Research Institute (IFPRI) [21–25], the experience of the Multi-Country Evaluation of the Integrated Management of Childhood Illnesses (IMCI) [17–19], and the literature on evaluation of complex interventions [26–30] to build the evaluation designs for Alive & Thrive. These experiences suggest that a strong impact evaluation strategy has five defining characteristics: (1) aligns with project objectives and activities; (2) addresses the question “What might have happened in the absence of the project?” and attempts to establish designs that best indicate the extent to which a causal link can be established between the project activities and the outcomes of interest; (3) follows the chronological and logical progression of the project cycle; (4) uses a variety of data collection methods to triangulate findings; and (5) is undertaken in a way that supports implementation objectives and does not create perverse incentives. To ensure that the design process for Alive & Thrive yielded an overall evaluation design that addressed these characteristics, we undertook the following steps, which are described in further detail below:

1. Identifying design options for attribution and intervention components to be evaluated;
2. Establishing valid comparison groups;
3. Choosing key impact indicators based on specific program goals;
4. Ensuring rigor in measurement and data collection (establishing conceptually grounded and solid data collection approaches, including sampling, questionnaire and survey design, and training and field supervision);
5. Documenting program impact pathways through process evaluation.

Step 1. Identifying design options for attribution and selecting intervention components to be evaluated

Attribution of impact to programs or interventions is a primary goal for any impact evaluation. However, it is not always possible to establish best-practice evaluation designs that enable attribution. The difference-in-differences (or double-difference) method allows the most reliable attribution of impact to the interventions [23, 25]. This method relies on comparing changes from baseline (using data collected before interventions are implemented) to follow-up (using data collected at endline) to develop a before–after comparison. Randomizing which groups get the interventions and which do not further strengthens the double-difference model and allows a probability design to be put in place [12].

In the case of Alive & Thrive, developing a design that incorporated both the double-difference approach and randomization required, first, that areas be randomly allocated to intervention and comparison areas. Second, it required that data be collected from areas covered by Alive & Thrive (the intervention group) as well as those not covered by Alive & Thrive (the comparison group) before and after Alive & Thrive interventions were implemented. As shown in figure 1, the difference between the change in outcomes for intervention (Alive & Thrive) and comparison (non-Alive & Thrive) groups in the baseline and endline surveys provides the difference-in-differences impact estimates. Even though there are other methods of attributing the impact of interventions [31], the double-difference approach with randomization allows a “clean” estimate of the impact attributable to the program by taking into account the changes in outcomes that may occur among the nonbeneficiary population as a result of nonprogram factors such as secular trends, other types of programs (e.g.,

![Illustration of the double-difference approach to capture change over time and differences between intervention and comparison groups. A&T, Alive & Thrive](image)

FIG. 1. Illustration of the double-difference approach to capture change over time and differences between intervention and comparison groups. A&T, Alive & Thrive
infrastructure), or climate-related or other types of shocks that may have independent positive or negative impacts on the study’s main outcomes.

A significant challenge for such designs in the context of relatively complex interventions, however, is identifying which components of the interventions can accommodate these types of evaluation designs. Alive & Thrive interventions include three main components to address several proximal and distal barriers to optimal IYCF practices in the three countries: systems strengthening to deliver community-based interventions such as IYCF counseling by health workers, mass media campaigns to shape social norms and support counseling interventions, and policy engagement and advocacy to address environmental factors such as country compliance related to the International Code of Marketing of Breast-Milk Substitutes. The evaluation portfolio for Alive & Thrive, therefore, needed to be tailored to the distinct needs of these three components as well as to the specific implementation plan and design in each of the three countries. Alive & Thrive’s interventions are delivered through different types of service delivery platforms [32–34]. At the core of the intervention portfolio for each country, however, is skilled counseling for IYCF by different types of service providers through standardized training of service providers, community- and mass-media activities to shape broader perspective, and policy and advocacy activities. The service provision platform in Bangladesh is a large-scale nongovernmental organization (NGO) volunteer platform; in Ethiopia, it is the government health extension platform; and in Vietnam, it is health workers at local health facilities. The mass media-based communications interventions in all three countries are delivered through a mix of radio, TV, and local video.

Not all of the above interventions and delivery platforms are easily amenable to the use of randomized evaluation methods (probability evaluations), which allow attribution of impact to the programs being evaluated. Thus, a critical decision point in the early design process for the Alive & Thrive interventions was to identify which of the interventions and delivery platforms would be selected for the primary impact evaluation (including attribution) in the three target countries. Within the Alive & Thrive portfolio, the mass media campaigns and policy advocacy were not fully amenable to differential allocation to intervention and nonintervention areas to create comparison groups that did not receive these interventions, and thus the evaluation designs to capture their impact could not use randomization. However, the interventions to strengthen the community-based delivery of behavior change communication (BCC) activities were the most amenable to a rigorous evaluation because they could potentially be randomized to different groups, and hence, were a focus for the impact evaluation design.

For the community-based interventions, therefore, we designed evaluations using quantitative and qualitative data collection methods to measure and attribute impact and to generate learning about pathways to impact. For the mass media interventions, we used data from the impact evaluation to assess the plausibility of impact. For the policy engagement and advocacy component, the evaluation approach drew on the nutrition and health policy process literature [35–37] and emerging innovations in policy advocacy research [38], and kept a focus on documenting shifts in the overall policy environments for nutrition and IYCF practices in each of the three countries. Finally, the costing component of the Alive & Thrive evaluation activities in each country drew on a well-accepted costing methodology [39–42], based both on the literature and on input from experts in costing nutrition interventions. An overarching feature of the design process was to maintain the emphasis on the “program theory” of the Alive & Thrive interventions, given the aim to unpack and document pathways to impact.

In the remainder of this paper, we lay out the details of the evaluation design considerations for the major intervention components of the community-based programs for IYCF counseling and the mass media interventions. The evaluation design for the policy and advocacy component of Alive & Thrive is described briefly in Hajeebhoy et al. [34], and the costing assessment methods will be described in subsequent papers. The process evaluation design process and evaluation designs are described in detail by Rawat et al. [43] in this Supplement.

Step 2. Establishing valid comparison groups

A key goal for the quantitative impact assessment of a program is to be able to compare outcomes for program beneficiaries with what these outcomes would have been had the program not been implemented at all. Alternatively, comparisons might be made between groups receiving different intervention packages. In Alive & Thrive, a central goal of the impact evaluation was to compare groups of children exposed to Alive & Thrive’s interventions with comparable groups of nonexposed children. In order to attribute and estimate the impact of Alive & Thrive’s program using a double-difference approach, as noted above, it was important that Alive & Thrive and non-Alike & Thrive groups be identical in all characteristics except for being exposed to Alive & Thrive, so that the only plausible explanation for differences in changes in stunting and IYCF practices between groups from baseline to endline would be the package of interventions. A major challenge when identifying an appropriate comparison group that is not exposed to Alive & Thrive is selection bias (areas chosen for coverage with Alive & Thrive areas could be fundamentally different from non-Alike & Thrive areas in ways that bias the outcomes and impacts). One option
for establishing an appropriate comparison group is to randomize communities or individuals to the interventions. In Alive & Thrive, the interventions were targeted at the entire community, and implementation was controlled at larger geographic units; randomization therefore had to be done at the geographic cluster level (areas), rather than the child level to be consistent with program implementation. This means that groups of children from different areas (rather than children within areas) would be compared when assessing differences between Alive & Thrive and non-Alive & Thrive areas, and that the children from the comparison areas would not receive program interventions for the whole duration for the project. When scaling up is a key goal of the program, withholding programming in comparison areas for the entire duration of the project may not be possible. One solution is to use a stepped-wedge approach to stage program rollout. Another approach is to set up comparison groups where certain program elements are absent or where interventions are delivered with differing intensities, and the choices of where to hold back program elements or to scale back on intensity are made randomly rather than purposefully. Where randomization is not feasible at all, options for establishing valid comparison groups include matching by design, or by analysis using econometric methods such as propensity score matching (PSM) or nearest neighbor matching (NNM) [21].

For Alive & Thrive, a key challenge was that the planned rapid scale-up would limit the number of areas that would remain uncovered by Alive & Thrive, thus eliminating the possibility of using these areas for comparison. Additionally, given the large number of geographic implementation areas, finding comparison groups for all implementation areas was not possible. Hence, in two of the final evaluation designs, described later, only a subset of implementation areas in each country was chosen as the target geographic area for the impact evaluation.

**Step 3. Choosing key impact indicators based on specific program goals**

Alive & Thrive’s overall goals are to improve exclusive breastfeeding and adequate complementary feeding practices and reduce childhood stunting. With these goals in mind, the following key impact indicators were identified:

- World Health Organization (WHO)-recommended indicators to measure IYCF practices [45]
  - Early initiation of breastfeeding
  - Exclusive breastfeeding under 6 months
  - Continued breastfeeding at 1 year
  - Introduction of solid, semisolid, or soft foods
  - Minimum dietary diversity
  - Minimum meal frequency
  - Minimum acceptable diet
- Consumption of iron-rich or iron-fortified foods
- Stunting (height-for-age < -2 z-scores)

In addition, in Bangladesh, where multiple micronutrient supplements are available for purchase nationally and are also being promoted by BRAC, the main implementer for the community-based counseling interventions, anemia was included as an indicator.

**Step 4. Ensuring rigor in measurement and data collection**

Ensuring that data collection is relevant and high quality requires attention to how impact indicators are measured, which age groups are included in the impact assessment measurements, and the sample sizes needed for impact estimates, as well as the overall relevance of the data collection to the interventions and the contexts. When measuring impacts on child IYCF and anthropometry, it is also important to consider the optimal ages for measurement of the different impact indicators.

**Children’s age for assessing impact of Alive & Thrive on stunting**

The assessment of the impact of nutritional interventions on child anthropometry (as well as other nutrition outcomes) should consider the age at which assessments should be made to detect the greatest difference between intervention and comparison areas. Evidence suggests that the earlier children are exposed to nutrition interventions, such as the promotion of improved IYCF practices and nutrition supplementation during their first 2 years of life, and the longer the exposure during this period, the more they are likely to benefit in terms of linear growth [44–48]. There is also evidence that nutrition interventions beyond the first 2 years have much less impact on stunting than those delivered before 2 years of age. Thus, although nutrition interventions should reach infants and young children during their first 2 years, the accrued impacts of these interventions should ideally be measured once this period of greatest potential benefit is concluded [22]. To illustrate this point, figure 2 presents the mean height-for-age z-scores (HAZ) of children in Bangladesh (2007) and shows that the main period of growth-faltering is when children are between 0 and about 18 to 20 months of age; changes after this age are minimal compared with the large drop in HAZ observed in the preceding period. For this reason, it is expected that if children are exposed to Alive & Thrive’s investments during the first 24 months of life, the greatest impacts on height (and stunting) will be observed once they have completed their full exposure to the intervention (i.e., after 24 months of age). Measuring Alive & Thrive’s impact on stunting at an earlier age (e.g., under 12 months) could underestimate the real impact of the intervention because children do not get...
the intervention during the entire period from 0 to 24 months of age. Thus, for Alive & Thrive, the impacts on stunting will be primarily assessed among children older than 24 months at endline.

**Age of impact assessment for IYCF practice outcomes**

The issue of age of impact assessment is also pertinent to the assessment of the practice-related outcomes of IYCF practices, as well as the use of health services in the critical age window up to 2 years of age. Specifically, certain WHO-recommended IYCF indicators [46] can only be assessed in certain age groups, while others are assessed in other age groups. Some of the main age groups in relation to a few of the major practice outcomes are shown in figure 3. The implications of this for the impact assessment are that sample sizes and sampling plans need to be adjusted to assess differences between Alive & Thrive and non-AIve & Thrive areas for practices that occur only within specific age windows. For example, children 0 to 5.9 months of age are sampled to evaluate impact on exclusive breastfeeding, children 6 to 23.9 months of age to evaluate impact on other IYCF practices and anemia, and children 24 months of age and older to evaluate impact on stunting. Sample size estimates for Alive & Thrive evaluations were therefore derived separately for each country, using Demographic and Health Survey data for estimated baseline prevalences of the three impact indicators (exclusive breastfeeding, dietary diversity in complementary feeding, and stunting). Effect sizes were derived from the literature on effectiveness of BCC interventions [7, 47]. Table 1 depicts the estimated sample sizes for the evaluation.

**Repeated cross-sectional surveys versus individual-level longitudinal studies**

A major consideration for impact evaluations targeting infants and young children for interventions is whether to use repeated cross-sectional surveys that measure different children (of the same age) in the same communities at baseline and endline, or whether to measure the same children over time. In program effectiveness evaluations, it is probably more appropriate
to use repeated cross-sectional surveys in the same communities to enable estimation of the overall public health impact of the program rather than the impact on specific individuals tracked over time. For evaluations that aim to measure impacts on stunting, our experience indicates that there is often a strong possibility that initially sampled or measured children will age out of the target under-two period for interventions before program implementation is robust enough to ensure adequate coverage, penetration, and utilization. Even if children are not yet 2 years of age when the program is fully operational, they are likely to be exposed to the program when they are too old and for too short a time to really benefit, resulting in an underestimation of the program impacts. Repeated cross-sectional surveys of children in the correct “impact age group” for the intervention are a more robust approach in these situations, as they best capture the impact among children who should have been exposed to the program prior to the evaluation, while also minimizing the risks of having to re-establish a baseline. Furthermore, repeated cross-sectional surveys of children in the same age group offer the flexibility of extending the evaluation endlines or aligning the evaluation data collection more easily with implementation timelines. A challenge with repeated cross-sectional surveys is the issue of seasonality, which might either falsely attenuate or amplify the program impacts seen, even if the intervention and comparison groups are equally affected by seasonality. This can be avoided by ensuring that the repeated cross-sectional surveys are done at the same season of the year. In the Alive & Thrive surveys, repeated cross-sectional surveys are used, and specific attention is paid to survey timing (data collection in the enumeration areas is matched to within 1 or 2 weeks at the same time of the year) to ensure minimal seasonal variation. Another challenge is that repeated cross-sectional surveys can require larger sample sizes, and thereby lead to higher costs, than child-level longitudinal studies where individual children serve as their own controls, lending more statistical power to the study. This higher cost of sampling a larger number of children may be offset by the logistical challenges of tracking down individual children over time in a longitudinal study.

**Ensuring rigor in measurement of impact indicators and influencing factors**

In developing the survey-based methods for measuring the impact of Alive & Thrive’s interventions, we aimed to carefully measure the impact indicators and to capture exposure to Alive & Thrive’s delivery platforms and interventions, as well as child, maternal, household, and community factors that might influence program impact.

**Impact indicators.** Although child stunting is the key indicator in Alive & Thrive, the surveys include measurements of child height and weight, which enable computation of all three primary indicators of anthropometric growth in children, i.e., stunting, underweight, and wasting. Standard measurement tools are used to measure height and weight, and the WHO (2007) reference standards for anthropometry are used to estimate z-scores and the prevalence of stunting, underweight, and wasting [48]. For IYCF practices, we drew on the core set of WHO infant and young child feeding indicators [46, 49], which have...

---

### TABLE 1. Sample size assumptions for Alive & Thrive impact evaluation designs and estimated sample sizes<sup>a</sup>

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Bangladesh (cluster-randomized probability design)</th>
<th>Vietnam (cluster-randomized probability design)</th>
<th>Ethiopia (before–after adequacy design)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline (estimated from national surveys)</td>
<td>Minimum detectable effect size (percentage points)</td>
<td>No. of clusters/enumeration areas</td>
</tr>
<tr>
<td>Stunting (24–48 mo)</td>
<td>43</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Exclusive breastfeeding</td>
<td>44</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Dietary diversity</td>
<td>50</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>Stunting (24–59 mo)</td>
<td>30</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>Exclusive breastfeeding</td>
<td>44</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>Dietary diversity</td>
<td>36</td>
<td>6.5</td>
<td>40</td>
</tr>
<tr>
<td>Stunting (24–59 mo)</td>
<td>49</td>
<td>8.0</td>
<td>75</td>
</tr>
<tr>
<td>Exclusive breastfeeding</td>
<td>33</td>
<td>6.5</td>
<td>75</td>
</tr>
<tr>
<td>Dietary diversity</td>
<td>46.5</td>
<td>5.5</td>
<td>75</td>
</tr>
</tbody>
</table>

<sup>a</sup> All sample size estimates were carried out using a power of 80%, an intracluster correlation (ICC) of 0.01, an alpha of 0.05, and a one-sided test (the assumption being that Alive & Thrive interventions will have some positive impact on the key impact indicators).
been recommended for widespread use to allow cross-country and cross-study comparisons. The impact on anemia is being assessed using field-friendly Hemocue machines, but only in Bangladesh, where there is an effort to support and promote the use of micronutrient powders containing iron.

*Exposure to Alive & Thrive interventions and delivery platforms.* Alive & Thrive's interventions are being delivered by different types of service delivery platforms in each of the countries. Thus, at baseline, we included a small set of exposure measures on the anticipated core implementation platforms. These exposure measures are being expanded in the process evaluation surveys as the details of the implementation platforms and implementation activities evolve through the implementation phases. Overall, utmost attention is needed to ensure that detailed data on exposure to intervention platforms are gathered to enable a deep understanding of program penetration and utilization. In the surveys that generate the data for the double-difference analyses, therefore, we are gathering adequate information on exposure to the community-based interventions as well as other Alive & Thrive interventions, such as media exposure and awareness and purchasing patterns of fortified complementary foods and other recommended foods.

*Child, maternal, and household factors.* There is a substantial literature that recognizes the influence of contextual factors within the household and community that might influence intervention uptake and impact [50–53]. However, few evaluations or research studies gather data on these factors. In Alive & Thrive, we use an expanded version of the UNICEF conceptual framework for nutrition [54] to ground the development of the survey instruments and to ensure that we have detailed information on the multiple factors that may influence the primary outcome measures (table 2). This provides opportunities for subgroup analysis related to intervention benefits, with the caveat that sample sizes for such secondary analyses might be limited.

*Implementation characteristics.* With the understanding that implementation characteristics, particularly the capacities and motivation of frontline workers, can affect the fidelity, quality, and impact of an intervention [55, 56], we include in each survey interviews with the Alive & Thrive-related frontline workers serving each of the communities in the impact evaluation survey areas (table 2).

**Step 5. Documenting program impact pathways through process evaluation**

Traditionally, impact evaluations have focused on answering *what* impact programs or interventions have, with less attention to *how* or *why* impacts are achieved or not achieved. To address Alive & Thrive's second major objective for MLE—*generate learning on how to achieve and replicate Alive & Thrive's impact*—it is crucial to open the “black box” of the portfolio of interventions. Our purposes in doing this for Alive & Thrive are primarily to generate information on program impact pathways, as well as to enable improvements in program performance through the timely sharing and discussion of results with program implementers. Details on the theory-driven approach used to develop the process evaluation for Alive & Thrive, design considerations, and the resulting process evaluation designs for Alive & Thrive are presented by Rawat et al. [31] in this Supplement.

**Balancing engagement with objectivity**

In all the evaluation design steps described above, the Alive & Thrive evaluation design process was carried out in ways that assured adequate engagement with the implementation teams to ensure attention to program evolution and needs. In addition, the design process also included explicit considerations to ensure clear objectivity of the evaluation team. The Alive & Thrive evaluation team engaged with the program and management teams in three activities specific to each country’s impact evaluation: defining the scope of the impact evaluation and identifying the possible options for establishing a counterfactual; obtaining feedback on data collection instruments; and sharing results of baseline and process evaluation results and obtaining feedback on analytic results to enable more grounded interpretation. Objectivity, which is particularly critical in an impact evaluation of an intervention or program that does not involve blinded assignments, was addressed in the following ways:

- Ensuring that final decisions on protocol design, data collection plans, and instruments lay with the evaluation team;
- *Independent data collection* by data firms that were not involved with the program implementation team;
- *Independent data analysis and primary interpretation of results* by the evaluation team.

In the next section, we describe the impact evaluation designs that were developed for the Alive & Thrive initiative in each of the focus countries.

**Results**

**Impact evaluation designs used in the three Alive & Thrive countries**

Randomizing the choice of Alive & Thrive program areas was feasible both in Bangladesh and in Vietnam because the geographic rollout and scaling-up plans were gradual. Nonetheless, the rapid program scaling-up plans meant that it was only possible to randomize
the interventions within a subset of implementation areas within the larger implementation area. This focus on a smaller implementation area led to a smaller number of clusters available for the randomization, but the overall approach was a good compromise that helped address the tradeoff between implementing at scale and establishing a solid evaluation design. In nonevaluation implementation areas, Alive & Thrive is

<table>
<thead>
<tr>
<th>Domain of data collection</th>
<th>Survey respondent</th>
<th>Baseline survey</th>
<th>Process evaluation survey—focus on implementation</th>
<th>Process evaluation survey—focus on utilization</th>
<th>Impact survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Household sample and identification</td>
<td>Head of household</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2. Household roster</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3. Maternal Health (pregnancy and postnatal care) and Family Planning</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4. Child immunizations and health</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5. IYCF practices</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6. IYCF knowledge, attitudes, and perceptions</td>
<td>Grandmother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>7. Child care arrangements</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8. Food security and diet diversity</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9. Socioeconomic status, employment, and economic events</td>
<td>Head of household</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>10. Social assistance and participation in programs</td>
<td>Head of household</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Market and facilities access</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>12. Information access</td>
<td>Respondent mother</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Woman’s empowerment</td>
<td>Respondent mother</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Water, sanitation, and hygiene</td>
<td>Mother of the index child</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Anthropometry and hemoglobin</td>
<td>Children in all impact indicator age groups</td>
<td>X</td>
<td>X (only IYCF age group)</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Mother of the index child</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frontline workers and implementers</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Demand for services, work load, and time commitments</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2. Exposure to training</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3. Motivation and job satisfaction, self-efficacy, and confidence</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4. Technical knowledge and skills: IYCF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5. Supervisory support</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6. Basic demographics</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
documenting scale and quality using monitoring and process evaluation data. Using this approach allows the program to go to scale rapidly without compromising on impact evaluation goals and methodological needs, thus balancing what are potentially competing demands.

In Ethiopia, no comparison areas were available because all areas within the larger geographic units, the woredas, were targeted simultaneously for implementation and scale-up; this precluded the use of plausibility design for the Ethiopia impact evaluation.

In all three countries, however, the impact evaluation strategy will ultimately capture the synergistic impact of Alive & Thrive’s community component along with mass media, compared with the mass media interventions alone.

The final evaluation designs are described below and are also summarized in Table 3.

TABLE 3. Impact evaluation objectives, design, samples and sample sizes in Alive & Thrive evaluations in Bangladesh, Vietnam and Ethiopia

<table>
<thead>
<tr>
<th>Evaluation component</th>
<th>Bangladesh</th>
<th>Vietnam</th>
<th>Ethiopia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective of main impact evaluation</td>
<td>To evaluate the combined effect of the Alive &amp; Thrive frontline worker and mass media interventions, compared with the mass media interventions alone, on IYCF practices and child undernutrition</td>
<td>To evaluate the combined effect of the Alive &amp; Thrive social franchise and mass media interventions, compared with the mass media interventions alone, on IYCF practices and child undernutrition</td>
<td>To estimate the change in IYCF practices and child undernutrition over time, in areas covered by Alive &amp; Thrive interventions delivered through multiple platforms</td>
</tr>
<tr>
<td>Impact evaluation design</td>
<td>Cluster-randomized (probability) evaluation design comparing Alive &amp; Thrive-intensive and -nonintensive interventions</td>
<td>Plausibility design for mass media interventions comparing exposed with non-exposed households</td>
<td>Cluster-randomized (probability) evaluation design for IYCF franchises</td>
</tr>
<tr>
<td>Geographic sample</td>
<td>20 subdistricts in 5 divisions, randomized to 10 subdistricts that are Alive &amp; Thrive-intensive and 10 that are Alive &amp; Thrive-nonintensive (media alone)</td>
<td>40 commune health centers from 4 provinces, 20 that receive Alive &amp; Thrive-intensive interventions and 20 that receive mass media alone</td>
<td>Before–after adequacy design for measuring community activities and mass media, with some possibility of exploiting natural variation in implementation rollout to establish high- and low-exposure areas for post hoc analyses</td>
</tr>
<tr>
<td>Sample size for baseline and endline surveys</td>
<td>4,400 children in the following age groups: 0–6 mo: 1,000 (500 intervention, 500 comparison) 6–23 mo: 1,000 (500 intervention, 500 comparison) 24–47.9 mo: 2,200 (1,100 intervention, 1,100 comparison)</td>
<td>4,000 children in the following age groups: 0–6 mo: 1,000 (500 intervention, 500 comparison) 6–23 mo: 1,000 (500 intervention, 500 comparison) 24–47.9 mo: 2,000 (1,000 intervention, 1,000 comparison)</td>
<td>4,000 children in the following age groups: 0–6 mo: 600 6–23 mo: 900 24–47.9 mo: 1,500</td>
</tr>
</tbody>
</table>

Bangladesh

In Bangladesh, implementation of Alive & Thrive interventions is done through BRAC, a highly reputable, large NGO with a nationwide reach and substantial prior engagement with program-focused research and evaluation. This choice of implementation partner also had positive implications for the design of the Alive & Thrive evaluations, and there were few obstacles to randomizing the intervention to geographic units within which the intervention implementation could be relatively well contained (in this case, subdistricts). In addition, the central oversight over field implementation within BRAC also ensured that there would be better control over the intervention and thus strengthened the possibility of enhanced fidelity to the intervention design.

In Bangladesh, therefore, the impact evaluation for Alive & Thrive’s community-based rural platform utilizes a cluster-randomized design in 20 rural
subdistricts that are part of BRAC’s Essential Health Care (EHC) program platform for Alive & Thrive. The impact evaluation compares subdistricts that receive the nationwide mass media campaign and the intensive in-person counseling intervention implemented by BRAC with subdistricts that are only exposed to the national mass media campaign. The baseline survey for the impact evaluation was completed at the end of July 2010. The endline survey will be conducted at exactly the same time of year in 2014 (see fig. 4 for the overall evaluation design).

**Vietnam**

In Vietnam, the intervention is linked with the government health services and entails implementation at the commune, district, and provincial levels. However, the knowledge that implementation is well controlled at the provincial level and the lack of comparability across provinces made it impossible to randomize the interventions across provincial levels (other provincial governments would probably not have appreciated their entire province being a “comparison” area). At the same time, in Vietnam it was feasible to have discussions with the provincial governments on setting up intervention and comparison communes within their provinces and about using randomization, which permitted the development of a double-difference, cluster-randomized evaluation design.

In Vietnam, the impact evaluation of the social franchise model put in place by Alive & Thrive is therefore based on a cluster-randomized evaluation design. Forty Commune Health Centers (CHCs) from four provinces were selected to be part of the evaluation area. CHCs were randomly assigned to either standard government service or Alive & Thrive IYCF franchise plus standard government service. The baseline survey for the impact evaluation of the franchise model was completed at the end of August 2010. The endline survey will be conducted at exactly the same time of year in 2014 (see fig. 5).

**Ethiopia**

In Ethiopia, IYCF counseling interventions were to be integrated and linked with the overall government-led health extension platform to ensure that intervention delivery was at large scale and sustained within a government system. In order to enable rapid integration of IYCF with the health system in this situation, Alive & Thrive worked closely with a third party bilateral partner that was already well linked with the Ethiopian health system and had training and supervisory systems already set up. Given the prior arrangements between this implementation partner and the Ethiopian health extension platform, and a mandate to operate at scale, it was not feasible to use a randomized, double-difference evaluation approach in this context.

---

**FIG. 4. Overall evaluation design for comparing Alive & Thrive-intensive and Alive & Thrive-non-intensive interventions in Bangladesh**

- **60 rural subdistricts**
- **20 (paired) rural subdistricts**
- **Randomized**
- **10 subdistricts**
  - A&T intensive
  - Intensive IYCF counseling by BRAC frontline workers + mass media
- **10 subdistricts**
  - A&T non-intensive
  - Standard care by BRAC frontline workers + mass media only
- **Baseline survey (April–July 2010) and early process evaluation (late 2010)**
- **Process evaluation survey on implementation (September–October 2011) and qualitative research**
- **Process evaluation survey of implementation and utilization (subsample only, June–July 2012) and qualitative research**
- **Process evaluation survey of implementation and utilization (all areas, April–July 2013)**
- **Endline survey (April–July 2014)**
This led to a substantially different impact evaluation design in Ethiopia, one that focuses on documenting the rollout of IYCF counseling interventions through the health extension platform and on estimating the adequacy of impact rather than on attributing impact to Alive & Thrive's specific interventions. In Ethiopia, therefore, the impact evaluation will capture the impact of Alive & Thrive's community component (within the Integrated Family Health Program [IFHP] components) along with any accompanying synergistic media communications. The impact evaluation has been designed to capture adequacy of change in Alive & Thrive indicators (stunting and IYCF practices). A before–after design, without a comparison group, was chosen because of the scale of operations of IFHP and the difficulty in this situation of identifying a valid comparison area. Furthermore, in Ethiopia, Alive & Thrive was expected to implement activities in four regions (Southern Nations, Nationalities, and People's Region [SNNPR]; Tigray; Amhara; and Oromia), and implementation of Alive & Thrive interventions was to be carried out in a phased manner with intervention launching first in Tigray and SNNPR. Therefore, the impact evaluation surveys are carried out in the first-phase regions (Tigray and SNNPR) in order to ensure maximum exposure to Alive & Thrive interventions in the period between the baseline and the endline surveys. The baseline survey was conducted in July 2010, and an endline survey will be done exactly 4 years later. The adequacy study design in Ethiopia, even more than in the other countries, will need to rely substantially on the process evaluation data and on robust epidemiological analyses to enhance the plausibility of results, including dose–response analysis and multivariate regression modeling. Figure 6 provides an overview of the Ethiopia evaluation design.

### Summary of country studies

**Table 3** summarizes the specific objectives, study designs, and sample sizes for each country. Detailed process evaluation plans and protocols were also developed for each country, using an "engaged but objective" process of interactions with the implementation teams for Alive & Thrive.

### Ethical approval

Data collection for the impact and process evaluation activities in Alive & Thrive received clearance from the institutional review board at IFPRI; country-specific data collection activities also received clearance from institutional review boards in Ethiopia (Addis Continental Institute of Public Health), Vietnam (Viet Nam Union of Science and Technology Association), and Bangladesh (Bangladesh Medical Research Council).

### Discussion

This paper describes the main principles used in the design of the Alive & Thrive impact evaluations, with a specific focus on bringing together best practice in program evaluation with the technical demands of a nutrition-focused evaluation and the complexities and realities of large-scale program implementation.
By providing an overview of the major technical considerations and the evaluation design process followed by the evaluation design team in this context, this paper also provides guidance to other researchers and program implementers for developing a high-quality, technically sound, yet contextually relevant and practically feasible evaluation design. In the context of Alive & Thrive, the balance of technical, implementation, and contextual factors together led to the development in Bangladesh and Vietnam of probability impact evaluation designs that allow clear attribution to the core intervention strategy, i.e., cluster-randomized pre–post evaluations, and in Ethiopia to an adequacy evaluation, which does not allow clear attribution but was more contextually appropriate. Importantly, with an embedded process evaluation and costing studies, the overall evaluation for Alive & Thrive has the potential to enable a solid, sustained legacy of evidence of impact and of processes through which impact was achieved through a data-driven, rather than only experiential, understanding of the types of factors that influence program impact in these contexts.

The development of thorough, yet relevant and feasible, program evaluations is not without its challenges. In the Alive & Thrive evaluation design process, an important challenge was the process of identifying the best possible comparison areas for Alive & Thrive interventions in the different contexts, even in situations where the implementation team was supportive of a full impact evaluation. In this case, the demands of holding back implementation in comparison areas challenged the mandate of “operating at scale.” This was overcome by choosing a subset of overall implementation areas that would be protected as “impact evaluation areas,” but this also meant that the evaluation design itself only covered a subset of the larger set of implementation areas. This second challenge is being addressed by gathering process evaluation data on implementation quality and service utilization in randomly chosen implementation areas that are outside the impact evaluation areas. Another challenge, this time in the case of Vietnam, is the issue of possible contamination of comparison commune health centers with Alive & Thrive training materials or actual training, since the overall intervention is implemented and coordinated by the provincial authorities. This is also being documented in the context of the process evaluation, to estimate the extent to which contamination might mask the true impact of the interventions.

Another learning from the Alive & Thrive evaluation process is that good evaluations demand flexibility and adaptability. In the context of the impact evaluation for Alive & Thrive, given the routine challenges of rolling out a fine-tuned complex set of interventions for BCC, implementation timelines shifted to be later than anticipated. This, in turn, had implications for the potential for children in the impact evaluation areas to be exposed to the interventions at the right time and for a long enough period of time to fully benefit from the interventions. The endline for the impact evaluation was therefore moved from an initial date of 2013 to 2014. The demand for flexibility and adaptability is even more relevant in the context of process evaluations, which are ideally linked and sequenced with implementation activities. Last, but not least, impact and process evaluations that are done well and that span multiple years require significant investments in funds and human resources; in the case of Alive & Thrive, the impact and process evaluation budget is close to 10% of the overall grant.

The evaluation design process followed in Alive & Thrive and the resulting evaluation designs have

---

*The survey covered 75 enumeration areas in 19 woredas from Tigray and 37 woredas from SNNPR

**FIG. 6. Adequacy evaluation design for Alive & Thrive in Ethiopia**
reinforced several insights about designing and implementing program evaluations from several other areas, notably from economics [15, 23], nutrition [22], public health [19], and health promotion [57, 58], and an emerging area of work on implementation and evaluation of complex interventions [26, 27]. Our experiences also bring to this literature on program evaluation some major technical insights based on the biology of child growth in developing countries and considerations about potential to benefit [59] that relate to evaluations of interventions to improve child nutrition. On the basis of the literature and our experience in designing these evaluations in diverse contexts for Alive & Thrive, we highlight the following aspects to consider in designing impact evaluations for nutrition interventions.

First, start with identifying the major impact indicators for the impact evaluation, and then work out the age groups that need to be included in evaluation surveys to ensure that impact on these indicators can in fact be achieved and estimated appropriately. Related to this, it is important to ensure adequate sample size for each of the major impact indicators and related age groups, rather than rely on “bucket” survey samples of children under 5 years of age, for instance.

Second, for child nutrition outcomes, it is critical to clearly differentiate which age groups will be targeted to receive the interventions, versus the age groups that need to be used to assess impact on the outcomes of those interventions. This is particularly important for interventions targeted within the first 1,000 days of life that aim to address stunting, because linear growth is actively happening during this period and therefore its cumulative effect on stunting is changing rapidly over time until approximately 24 months of age. Thus, in evaluating impact on stunting, especially for designs based on two cross-sectional surveys, the target age groups for evaluation should be children 2 years of age or older who have (ideally) been fully exposed to the intervention and who are beyond the key period of rapid linear growth-faltering.

Third, start evaluation design processes very early in the project cycle and work in close collaboration with implementation teams, with the explicit goal of identifying the best possible model for establishing a counterfactual (i.e., establishing a valid comparison group, where possible through randomization), and work closely with implementation teams to explore all possibilities. In our experience, establishing randomized designs was not challenged by either unwillingness or perceptions of lack of feasibility in either Bangladesh or Vietnam; it simply required explicit discussions to this end early in the evaluation design process. In Ethiopia, randomization was constrained because of the ultimate modalities of implementation, which precluded even the establishment of a quasiexperimental, nonrandomized design.

Fourth, an investment in process evaluation will have wide-ranging benefits to implementers and evaluators. This requires careful attention to program theory and enough engagement with implementation teams during the process of program implementation and in the design of a detailed program impact pathway. It also requires, as has been discussed in the literature before, a cross-disciplinary orientation; the Alive & Thrive process evaluations draw on literature from nutrition, health systems, organizational theory, economics, sociology, and anthropology, to name a few.

Finally, staying theory-driven and conceptually grounded both in design and in measurement is central to enabling effective interpretations. Alive & Thrive evaluation instruments are grounded in the specific program theory of each of the intervention components, so as to capture and document access to Alive & Thrive interventions, as well as to measure and capture the contextual factors at the level of frontline workers, mothers, households, and communities that can influence the effectiveness of the interventions.

Conclusions

The Alive & Thrive evaluations were crafted with attention to many of the technical elements required to establish good program evaluations, while also balancing considerations related to implementation and contextual characteristics. The evaluation design process required evaluation training and experience in nutrition evaluations, as well as early and adequate engagement with implementation teams. In agreement with the call-to-action of a recent Center for Global Development review on evaluation [10], we conclude that establishing solid evaluation designs within large-scale programs is highly desirable and indeed, feasible. We caution that evaluation designs that are well aligned with program goals and local contexts require early engagement with implementers, substantial financial and human resources, and an overarching interest in establishing good evaluation designs.

Acknowledgments

We gratefully acknowledge the contributions of Disha Ali, Phuong Hong Nguyen, and Kuntal K. Saha, IFPRI research fellows and Alive & Thrive MLE coordinators in Ethiopia, Vietnam, and Bangladesh, respectively, to the design and implementation of country-specific evaluation activities. We are grateful for the inputs to the evaluation design process from key Alive & Thrive implementation partners in Bangladesh (Tina Sanghvi, Md. Raisul Haque, Faruque Ahmed, and Saisa Siraj), in Ethiopia (Teweldebrhan Hailu Abraha), in Vietnam (Nemat Hajeebhoy), and at Alive & Thrive headquarters (Jean Baker, Luann Martin, and Karin Lapping).
We thank Jean-Pierre Habicht and Gretel Pelto (Cornell University) and Jef Leroy (IFPRI) for critical reviews of earlier drafts of this paper.

P. Menon and R. Rawat led the design of the Alive & Thrive impact evaluations in all three countries and developed the overall approach for the theory-driven process evaluations. M. Ruel supported the overall evaluation design process and critical impact evaluation decisions. P. Menon drafted the first version of this manuscript. R. Rawat and M. Ruel critically reviewed and revised the manuscript. All authors read and approved the final submission.

The authors declare that there are no conflicts of interest.

Reference


