Food and Agriculture Organization
of the United Nations

Impact evaluation of HOME-GROWN SCHOOL FEEDING PROGRAMMES

Methodological guidelines

Impact evaluation of HOME-GROWN SCHOOL FEEDING PROGRAMMES

Methodological guidelines

Sara Giunti
University of Milan-Bicocca

Elisabetta Aurino
Imperial College London

Edoardo Masset
Centre of Excellence in Development Impact and Learning, London International Development Centre

and

Ervin Prifti
formerly Food and Agriculture Organization of the United Nations

Food and Agriculture Organization of the United Nations
Rome, 2022

Required citation:
Giunti, S., Aurino, E., Masset, E. and Prifti, E. 2022. *Impact evaluation of home-grown school feeding programmes – Methodological guidelines*. Rome. FAO. https://doi.org/10.4060/cb8970en

ISBN 978-92-5-135886-3

Cover photograph: Shutterstock

Contents

Tables

Figures

Boxes

Acknowledgements

The contributions regarding qualitative research methods and on sampling of research sites and informants for qualitative research were written by Pamela Pozarny, Senior Rural Sociologist at the Food and Agriculture Organization of the United Nations (FAO), and Zahrah Nesbitt-Ahmed, to whom we are indebted.

The authors wish to thank the FAO team for their invaluable support and inputs, especially, Cristina Scarpocchi, Florence Tartanac, Silvio Daidone, Luana Swensson, Alejandro Grinspun and Mari Kangasniemi, as well as Fabio Veras, research coordinator at the International Policy Centre for Inclusive Growth (IPC-IG), Brazil.

The authors sincerely thank to Donatella Marchi / studio Pietro Bartoleschi for the graphic design. Acknowledgements are also extended to Chiara Deligia (FAO) for her communication support.

Abbreviations and acronyms

ATE	average treatment effect
ATT	average treatment effect on the treated
ATU	average treatment effect on the untreated
BAZ	weight-for-age z-scores
CASU	Conservation Agriculture Scale Up (Zambia)
CCT	controlled clinical trial
DiD	difference-in-differences method
EMIS	Management Information System
FAO	Food and Agriculture Organization of the United Nations
FO	farmer organization
FGD	focus group discussion
FWER	family wise error rate
GSFP	Government of Ghana School Feeding Programme
HAZ	height for age z-scores
HGSF	home-grown school feeding programme
ITS	Interrupted Time Series
KII	Key Informant Interview (KII)
LATE	local average treatment effect
LMIC	low and lower-middle-income countries
MDE	minimum detectable effect
MOE	Ministries of Education
MTP	multiple testing procedures
NGO	non-governmental organization
NMK	Njaa Marufuku Kenya Project
OPM	Oxford Policy Management
P4P	Purchase for Progress
PAA	Purchase from Africans for Africa
PRONAE	Projecto de Alimentação Escolar
PSM	propensity score matching

PSU	primary sampling unit
PtoP	Protection to Production
RCT	randomized control trial
RDD	Regression Discontinuity Design
SDG	Sustainable Development Goal
SFP	school feeding programme
SSU	secondary sampling units
WFP	World Food Programme

USD	United States Dollars

List of countries

(Names in brackets are used in the text)

The Plurinational State of Bolivia (the Plurinational State of Bolivia)
The Republic of Botswana (Botswana)
Burkina Faso
The Federative Republic of Brazil (Brazil)
the Republic of Cabo Verde (Cabo Verde)
The Republic of Chile (Chile)
The People's Republic of China (China)
The Republic of Colombia (Colombia)
The Republic of Côte d'Ivoire (Cote d'Ivoire)
The Republic of Ghana (Ghana)
The Republic of Guatemala (Guatemala)
The Republic of India (India)
The Republic of Italy (Italy)
The Republic of Kenya (Kenya)
The Republic of Mali (Mali)
The Federal Republic of (Nigeria)
The Republic of Mozambique (Mozambique)
The Republic of Paraguay (Paraguay)
The Togolese Republic (Togo)
United Kingdom of Great Britain and Northern Ireland (United Kingdom of Great Britain and Northern Ireland)
United States of America (United States of America)
The Republic of Zambia (Zambia)

Introduction and purpose of this guidance note

School feeding programmes (SFP) are among the most common forms of social protection, reaching about 368 million children daily, for a global investment of United States Dollars (USD) 70 billion per year (World Food Programme [WFP], 2013). SFPs, including school meals or food rations to take home, aim to increase human capital investments in school-aged children by improving educational outcomes such as school attendance, completion and learning, and enhancing health and nutritional status. Home-grown school feeding (HGSF) initiatives stand out for linking SFP to agriculture development using food that is produced and purchased within the country. The overall goal of linking SFP to agriculture development – particularly to local small-scale production – is to reduce rural poverty by developing markets, generate a regular and reliable source of income for smallholder farmers, and provide support to overcoming barriers that prevent farmers from enhancing productivity.

Globally, many countries such as Brazil, Ghana, Kenya and Nigeria have undertaken HGSF (Drake *et al.*, 2016). However, with a few exceptions, there is a dearth of empirical evidence on the effectiveness and economic sustainability of such programmes with regards to the goals of enhancing farmers' incomes, food security and productivity. This evidence gap hampers the promotion of what has been framed as a potentially "win–win" intervention. Further, it hinders the scale-up of HGSF in the broader framework set by Sustainable Development Goal (SDG) 2, which aims to achieve food security and nutrition targets by promoting sustainable agriculture (FAO and WFP, 2018).

Since HGSF programmes are by definition cross-sectoral, with goals that span social protection, education, nutrition and agriculture, and involve two main beneficiary groups (school-age children and farmers), rigorous impact evaluations need to capture both education and the effects of nutrition on schoolchildren on the one hand, and agricultural impacts on farmers on the other (FAO and WFP, 2018). Given this complex nature, evaluating HGSF presents several methodological challenges. This note seeks to support practitioners by providing methodological guidelines for conducting rigorous impact assessments of HGSF programmes. It presents an overview of the main technical issues to be addressed depending on the characteristics of the context and of the intervention itself. While these guidelines are mainly designed for monitoring and evaluation (M&E) officers working for United Nations agencies, local governments or non-governmental organizations (NGOs), its contents can be of

interest to a wider audience of policymakers, researchers and practitioners interested in multi-sectoral, complex programmes linking agriculture and nutrition.

In this note, we mostly focus on the agricultural goals, as this is the area where the largest knowledge gaps remain (Gelli *et al.*, 2016; Sumberg and Sabates-Wheeler, 2011). Although we provide general indications for evaluating programme impact on all beneficiary groups involved, including school-going children, we mostly focus on the methodological challenges related to the estimation of the effects of HGSF interventions on farmers. Specifically, for these guidelines, we emphasize practical differences in evaluation focusing on two main food procurement modalities: a decentralized model where each school procures food from (smallholder) producers living in school catchment areas, and a more centralized model in which procurement occurs centrally or at district level.[1, 2] These are very common operating models for school food procurement. However, in practice, HGSF programmes can be implemented in many ways, and food procurement systems that combine elements from both schemes are frequently encountered.

The present guidelines provide practical answers to the following overarching questions:

- What is the rationale behind HGSF programmes? What are the main challenges in designing rigorous evaluations for HGSF programmes? What are the differences, when performing HGSF evaluations, between decentralized and centralized food procurement models?
- What is the theory of change behind HGSF? How do school meals and public food procurement affect beneficiaries in terms of nutrition and education for school children, and in terms of farm production, agricultural profits, and increased income for farmers? What is the role of "supporting factors" and of "contexts or structural mechanisms" in understanding the success or failure of the programmes?
- How can an adequate research design be selected to conduct an impact evaluation of HGSF? What are the most common experimental and non-experimental evaluation designs that can be implemented? How can these techniques be adapted to specific characteristics and to the food procurement model adopted by the HGSF programme?
- What are the benefits of implementing a mixed method impact evaluation? How does qualitative analysis enhance quantitative findings and strengthen impact evaluation findings?
- What main principles, approaches, and methods are employed when undertaking qualitative analysis as part of a mixed methods impact evaluation?
- What are the most suitable sampling strategies in face of different food procurement models?
- What are the outcomes to be measured to assess programme effects?
- What are the implications of external validity for the results of the evaluation?

---

[1] In semi-decentralized procurement schemes, both the funds to purchase the food and the procurement authority are transferred to intermediate structures, e.g. an NGO, as in Togo; catering companies, as in Ghana; or central kitchens, as in Tunisia, that are in charge of procuring and delivering food to schools. See Annex 7 in (FAO and WFP, 2018).

[2] In public administration literature, centralization refers to the "central government" and is used in the opposite sense of decentralized systems that can be at the regional, municipal, or, in the case of school feeding, at school level. The World Bank defines "decentralization" as the "transfer of authority and responsibility for public functions from the central government to intermediate and local governments or quasi-independent government organizations and/or the private sector."

Key messages

- From an agricultural perspective, HGSF impact evaluations aim to assess: whether HGSF generates additional food demand in the market; whether local farmers respond to the additional demand by producing more food and generating higher revenues; and whether a more stable demand and potentially higher revenues allow local farmers to overcome the barriers that prevent them from expanding their production and increasing productivity. Data collected to answer these questions should include: data on farmer household income and expenditure; purchase and sale prices for local foods; access to markets (including credit) and identification of intermediary transactions; production modalities and investments; labour and attitude to risk.
- If possible, running a feasibility study through qualitative methods is a good preliminary step to setting up the intervention and related evaluation design. Such analysis would shed light on the structure of local agricultural markets (e.g. number and size of local smallholders, productivity and production capacity, presence of cooperatives, etc.), identifying all stakeholders potentially affected by the programme, as well as the existing institutional capacity to deliver and uptake the programme. These elements should guide the design of the food procurement system that is most suitable for the local context. Furthermore, information gathered by qualitative data collection at this preliminary stage can help identify the mechanisms through which the programme would affect beneficiaries providing helpful insights for designing quantitative survey tools (e.g. include specific sections for measuring these mechanisms in survey questionnaires). A pilot of the intervention is also recommended to assess the acceptability of the programme among school children, farmers and other relevant stakeholders.
- The evaluator should have a good understanding of intervention design, i.e. the procedures to follow in selecting intervention areas, the selection criteria, which is important when selecting the most appropriate methodology for the impact evaluation.
- In order to guarantee an ideal counterfactual, treatment and comparison, groups should fulfil the following properties: average characteristics of the two groups must be identical in the absence of the programme; treatment should not affect the comparison group either directly or indirectly; outcomes of units in the control group should change the same way as outcomes in the treatment group, if both groups were given the programme (or not).
- In general, the larger the sample size, the more likely it is to be representative of the population from which the sample is taken. However, notable cost and time trade-offs are linked to sample size. Power calculations determine the minimum sample size sufficient for detecting statistically significant intervention effects. Power calculations have to be conducted separately for each outcome variable and the largest required

sample size among power calculations will be the final sample size. Power calculations need to be independently checked by practitioners skilled in statistics. In the context of HGSF programmes, an additional problem arises when constructing a representative sample of local producers, the population of farmers that fulfil the criteria to sell food to school feeding, e.g. quality standards, minimum level of guaranteed production, etc. may represent a small fraction of the entire population of local farmers. Oversampling the population subgroup eligible for the intervention allows for ensuring that targeted groups are included in the sample, even when representing a minority of the overall population of farmers.

- Qualitative methods can also ease the interpretation of quantitative findings. For instance, qualitative interviews or focus groups with the stakeholders can be conducted to identify the effects of spillover on population groups that have not been targeted for the intervention or to measure the non-economic outcomes of the programme. For qualitative analysis - a combination of randomization with purposeful sampling is one effective approach to sampling. This enables a level of confidence, which is indicated by recurrent findings and results through systematic random informant selection in varied locations, while also addressing unique cases, outliers, and comparisons with non-treated populations.
- HGSF interventions are implemented using a variety of modalities and specific local features and often simple generalizations of the results, from one context to the next, may not be possible. Evaluators and policymakers should not rely on acritical extrapolation of findings to new settings, rather they should identify the mechanisms that made the intervention work, or fail to work. For instance, they should consider the circumstances under which HGSF generates additional food demand in the market, whether farmers respond to the additional demand by producing more food, and whether this translates into a larger and more stable demand for food.

Background

A. RATIONALE BEHIND HOME-GROWN SCHOOL FEEDING PROGRAMMES

School Feeding Programmes provide meals, snacks or take-home rations to children, conditional on school enrolment and regular attendance. SFP aim to enhance child nutrition, improve school enrolment and reduce absenteeism. Programmes can also contribute to improved learning outcomes, by alleviating short-term hunger, enhancing cognitive abilities and increasing time spent in school (Drake *et al.*, 2017; Grantham-McGregor, Chang and Walker, 1998). The educational and nutritional effects of SFP can be boosted by complementary actions, e.g. deworming or providing micronutrients in the meals (Bundy, *et al.*, 2009). As a social protection intervention, school feeding can protect children's health and education from the detrimental effects of large shocks, such as droughts or conflicts (Singh, Park and Dercon, 2014; Tranchant *et al.*, 2018). By supporting accumulation of human capital, i.e. the stock of skills and health of individuals, SFP can have a life-long impact on increased earnings, employability and health at later stages (Burde *et al.*, 2015). Through this human capital channel, school feeding thus contributes to economic growth and to the achievement of the SDGs related to health, education and poverty.

Almost every country in the world, in some way and at some scale, provides food to its school-children (Gelli *et al.*, 2016). In high and upper-middle-income countries, SFPs are well integrated in the national education systems, and programmes are usually run by Ministries of Education (MOE). In low and lower-middle-income countries (LMIC), some kind of school feeding is generally available to children, although access is not always universal. Programmes are often funded by national and international donors and implemented by development or humanitarian organizations (FAO and WFP, 2018). However, in recent years there has been a shift in the financing and implementation of SFPs in LMIC towards increased ownership by national and local governments (Bundy, *et al.*, 2009; FAO, 2014; Swensson and Klug, 2017).

In the last decade, the idea of using SFPs as a vehicle for local agricultural development has received mounting attention.[3] Smallholder agriculture is a key source of income and food security for most poor people in LMIC. However, smallholder farming remains a low-return and rather risky activity (Davis *et al.*, 2017; Poulton *et al.*, 2006). In 2003 the New Partnership for Africa's Development (NEPAD) defined HGSF as a strategic initiative for promoting food security and rural development in Africa. Since then, HGSF programmes have been increasingly identified as an opportunity to improve the livelihoods of smallholder farmers and local communities by strengthening the nexus among nutrition, agriculture and social protection (Sumberg and Sabates-Wheeler, 2011; FAO and WFP, 2018). According to the HGSF rationale, households, small farmers and local businesses can benefit from the increasing and regular demand for food products triggered by SFP, through the removal of the structural constraints that smallholder farmers face to access capital (land, water, labour, technology, etc.), markets, information and credit.[4] For HGSF to work, school food procurement needs to be designed in a way that increases farmers' productivity and their ability to access the school food market (WFP, 2017). The innovative element of HGSF comprises supporting smallholder farmers by giving them access to a predictable and stable local market, such as that offered by school feeding, which is usually delivered for around 200 school days per year. This sustained market access in turn reduces overall uncertainty, and should make it easier for farmers to invest in more intensive and diversified food production and supporting activities (Kretschmer *et al.*, 2014). At the same time, the programme needs to incorporate specific tools to address the structural constraints on farmers' productivity. HGSF programmes can provide multiple forms of assistance to help farmers enhance productivity, e.g. storage solutions, support to forming farmer organizations, access to credit and training, to name a few examples.

[3] The 2008 food, fuel and financial crises stressed the role of school feeding programmes both as a social safety net for children living in poverty and food insecurity, and as a tool for stimulating local agricultural production and economic opportunities in rural communities (Lawson, 2012).

[4] Collier and Dercon (2009), however, criticise this view by arguing for "a more open-minded approach to different modes of production" as a way out of poverty in Africa.

B. IMPACT OF SCHOOL FEEDING PROGRAMMES: EXISTING EMPIRICAL EVIDENCE

A quantitative impact evaluation is an assessment of how an intervention or a policy affects some pre-selected outcomes. In this note, we mostly focus on quantitative impact evaluations, although we also discuss the role of qualitative methods at the design and interpretation phases. Statistical and econometric techniques can be implemented to rule out the possibility that any factors, other than the programme itself, may explain the observed impact. Experimental designs (e.g. randomized control trials) and quasi-experimental designs (e.g. quasi-natural experiments, difference-in-differences, regression discontinuity, propensity score matching) are usually considered as the most rigorous impact evaluation strategies from the viewpoint of quantitative methodology. Thus, studies exploiting these techniques can produce rigorous empirical evidence of the impacts of the intervention.

BOX 1. DEFINITION OF RIGOROUS EMPIRICAL EVIDENCE

Effects on child nutrition

Extensive empirical evidence on the impact of SFPs on children's growth and weight gain shows mixed results. Overall, SFPs lead to positive weight gains, while effects on linear growth are more mixed and often limited to younger children (Kristjansson *et al.*, 2007). A recent study in Ghana found that the local programme contributed to growth among girls in mid-childhood and the poorest children (Gelli *et al.*, 2019). Mixed findings are determined by multiple factors, including differences in programme objectives, modalities (e.g. school meals versus take home rations), and implementation (Jomaa *et al.*, 2011). However, the nutritional benefits of SFPs on energy, protein, and micronutrient intakes were documented in many different contexts (Afridi, 2010; Arsenault *et al.*, 2009; van Stuijvenberg *et al.*, 2001).

Effects on child education

Extensive evidence documents that SFPs increase school enrolment and attendance in different settings, with larger effects on the most disadvantaged groups in terms of access to education. Kazianga *et al.* (2009) found that both take-home rations and SFP interventions in Burkina Faso had a positive and statistically significant impact on the overall enrolment, especially girls. Afridi (2011) examined the effects of a school feeding intervention on enrolment and attendance in Madhya Pradesh (India). An increase was reported of 10.5 percent in girls' attendance in grade 1 in treated schools, while attendance for boys showed a positive but insignificant increase. Girls from scheduled tribes were marginally more likely to enrol due to SFP. During the recent conflict in Mali, emergency school feeding led to increased in enrolment by 11 percentage points and to about an additional half-year of completed schooling (Aurino *et al.*, 2018).

Source: Authors' own elaboration.

SFPs may have a positive effect also on cognition and learning achievements, particularly if supported by complementary actions such as de-worming and micronutrient fortification or supplementation (Bundy *et al.*, 2009). However, evidence of SFPs impacting cognitive ability is more nuanced than in the case of educational access, also because of variations in programme implementation modalities and selected target groups for existing evaluations (Lawson, 2012). Ghana's school feeding programme (GSFP) led to a moderate increase in test scores for the average pupil in treated schools, ranging between 0.12 and 0.16 standard deviations, with more remarkable learning and cognitive gains for girls, poorest children, and children from the country's most disadvantaged regions (Aurino *et al.*, 2018). In India, the local "Midday-meal" programme led to increases in reading and maths (Chakraborty and Jayaraman, 2016).

QUICK REFERENCES

The following reviews provide a good overview of the impacts of school feeding on child education and nutrition:

- Drake, L., Fernandes, M., Aurino, E., Kiamba, J. and Giyose, B. School feeding programmes in middle childhood and adolescence. *In Disease control priorities (third edition): Volume 8, Child and adolescent health and development*, edited by D. Bundy, N. de Silva, S. Horton, D.T. Jamison, G. Patton. Washington, DC, World Bank. **http://dcp-3.org/chapter/2428/school-feeding**
- Snilstveit, B., Stevenson, J., Menon, R., Phillips D. and Gallagher, E. 2016. *The impact of education programmes on learning and school participation in low- and middle-income countries, International Initiative for Impact Evaluation (3ie)*, Systematic review summary 7. **http://www.3ieimpact.org/evidence-hub/publications/systematic-review-summaries/impact-education-programmes-learning-school-participation-low-and-middle-income-countries**
- Kristajnsson, E.A., Gelli, A., Welch, V., Greenhalgh, T., Liberato, S., Francis, D. and Espejo, F. 2016. Costs, and cost-outcome of school feeding programmes and feeding programmes for young children. Evidence and recommendations, *International Journal of Educational Development*, 48, p.79-83. **https://www.sciencedirect.com/science/article/pii/S0738059315300134**

Effects on agriculture

The relevance of local purchases in stimulating the local economy and improving the food security of producers and their households has been qualitatively documented in several high and middle-income countries, i.e. Brazil, China, the United Kingdom of Great Britain and Northern Ireland, etc. (WFP, 2017). In fact, in settings such as the United States of America or the United Kingdom of Great Britain and Northern Ireland, SFPs started to fulfil primarily agricultural goals (Drake *et al.*, 2017). Innovative approaches to HGSF have been successfully tested and implemented in various country contexts at different stages of the programming and implementation cycle. The experience of Brazil shows how the regulatory framework can be shaped to channel a share of overall procurement requirements to smallholder farmers. More recent interventions in Brazil, Côte d'Ivoire, Ghana, Kenya (Njaa Marufuku

Kenya – Njaa Marufuku Kenya Project [NMK], "Eradicate Hunger in Kenya"), and Mali have been designed to stimulate agricultural outcomes as a strategic priority.

However, quantitative empirical evidence documenting the direct and indirect impacts of HGSF interventions on agricultural development and local communities is limited. So far, only a few isolated attempts have been conducted to estimate the effects of HGSF on smallholder farmers and local communities. Upton *et al.* (2012) exploited a natural experiment in which some schools in Burkina Faso received food imported from the United States of America while others from locally procured sources to compare delivery times and commodity costs. Cost savings of 20 percent were reported by agencies purchasing the food from local producers, while they were still meeting the government standards for food quality. Importantly, procurement of commodities from local producers did not distort market prices, and producers were able to realize higher prices since they knew there would be demand for their products throughout the school year (Upton *et al.*, 2012).[5] Additional evidence of the positive impacts of school feeding on local producers come from Guatemala, where sourcing school meals has shifted from industrial suppliers to local producers. Parents of school children participate in food supply and distribution, which has generated extra income for their households (Gockel *et al.*, 2009). During Indonesia's economic crisis in the 1990s, the government established that all food for SFP should be produced locally. Meals were prepared by local women's associations, and farmers reported that the project had increased their sales (Devereux, 2015). This limited evidence seems to support the idea that creating synergies between school feeding initiatives and agricultural procurement can increase revenues and incomes in the local agricultural sector (Lawson, 2012). A randomized control trial (RCT) is under way to analyse the impact of the national HGSF programme in Ghana (Gelli *et al.*, 2016).

C. FOOD PROCUREMENT MODELS

The rationale behind HGSF programmes draws on the idea that the local agricultural sector and the livelihoods of family farmers can be transformed for the better by the increased demand for food from local schools (Gelli *et al.*, 2016). While there is consensus among international agencies and funders, national governments, academics and practitioners about the potential of HGSF programmes to foster positive linkages or synergies between social protection and agricultural development, there is much less clarity and agreement about the *scale* at which the linkage between school feeding and agricultural development should take place.

In the context of HGSF planning, the term "local" is used interchangeably to indicate linkages with both small- and medium-sized food producers and different scale traders at the retail and wholesale levels (Sumberg and Sabates-Wheeler, 2011; FAO and WFP, 2018). Based on the context, governments

[5] The number of producers involved in food procurement varied based on commodity. Four large unions, composed of between about 600 and 5 000 individual members each, were identified to supply millet. Furthermore, vitamin-A fortified vegetable oil was purchased from a unique provider selected after a competitive tender. For additional details, see Upton *et al.*, (2012).

or international agencies can design institutional procurement schemes that support different types of farm holders linking the supply chain to community/district, regional or national markets (FAO and WFP, 2018). The implications of HGSF for agriculture development are strictly related to the structure of the food supply chain. The design of the impact evaluation should take into consideration the characteristics of the food procurement system when choosing the most appropriate research design, sampling strategies and data collection techniques for the HGSF programme in hand.

BOX 2. **DEFINITION OF FOOD PROCUREMENT**

With the term "food procurement," practitioners refer to all operations concerning food sourcing, buying and receipt of products. The scope of the entire system is to supply quality food to the SFP throughout the school year (Drake *et al*., 2016). Key activities linking food producers and school children include production, trade, transport, preparation and distribution (Gelli *et al*., 2012). Depending on the procurement model adopted, this translates into regulatory frameworks that regulate either direct links with smallholder farmers or interactions with traders operating as intermediaries in the market.

For an extensive description of public food procurement models see: FAO (2018), Strengthening sector policies for better food security and nutrition results: public food procurement, **http://www.fao.org/policy-support/resources/resources-details/en/c/1175509/**

Source: Authors' own elaboration.

Gelli *et al*. (2012) proposed a classification of supply chain models for HGSF programmes in three dimensions: key activities in the supply chain; level of activity; and actors in the supply chain. All steps linking food producers to schoolchildren are referred to as "key activities." "Level of activity" refers to whether the single supply chain element is occurring at the school, district, regional, national or international levels. The actors are all stakeholders involved in the various stages, e.g. smallholders and other producers, traders, intermediaries, teachers, parent-teacher associations, caterers, service providers such as school feeding coordinating agencies, communities and local institutions at different levels. On the other hand, Sumberg and Sabates-Wheeler (2011) distinguish HGSF procurement models on two "spatial" variables: the degree to which producers are clustered, and the proximity of producers to the point of consumption.

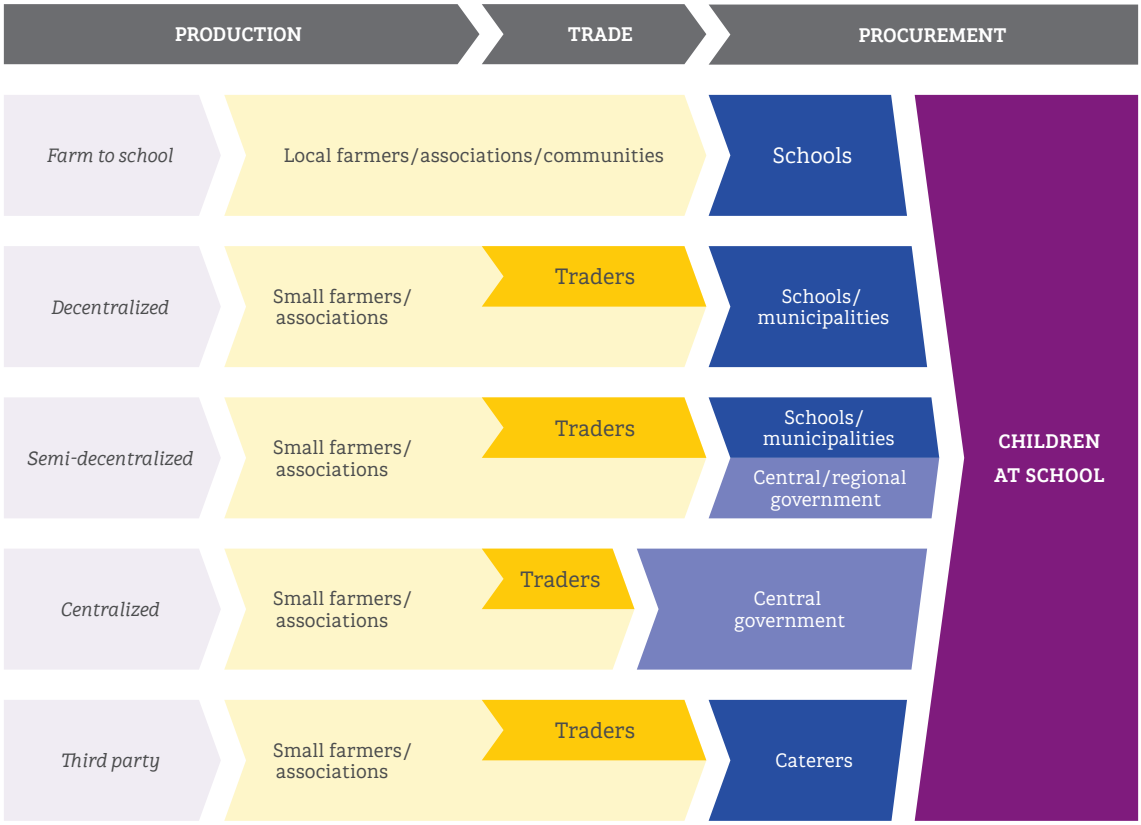
Central elements of distinction are also the level at which decision-making occurs (**centralization**) and whether activities are performed in-house or by a third party (**outsourcing**) (Gelli *et al*., 2012). Processing and service delivery can be either outsourced or decentralized to the community level. In the outsourced model, contracted service providers organize food procurement, arrange transportation, storage, preparation and distribution of meals to schoolchildren. The government usually provides funds based on a fixed payment per meal served (usually determined on enrolment/attendance figures).

The main advantages of centralized models are improved quality control and efficiencies through economies of scale, which are partly offset by the cost of transporting food from centralized warehouses to schools. In a decentralized model instead, transportation costs are limited, while quality control

and storage are the main difficulties, linked to guaranteeing a steady supply of food, food quality and safety. There is a lower risk of corruption with decentralized procurement, fewer bureaucratic hurdles and negligent delivery than with the more centralized (Gelli *et al.*, 2012). Also, approaches to decentralized procurement may integrate *ad hoc* supply-side supporting activities that boost the productivity of local farmers (Drake *et al.*, 2016). Additional models fall between these two extremes, e.g. partially decentralized models, or integrated farm-to-school models. Examples of procurement systems are presented in Figure1.

In practice, different organizational schemes can coexist and interact within the same project or geographical area. Frequently, in order to guarantee a school food menu throughout the school year, a combination of food procurement modalities operating at different levels of aggregation is necessary (Gelli *et al.*, 2012). For instance, it is common to find models that supply staples at a centralized level and fruits and vegetables at a decentralized (Drake *et al.*, 2016). See the case study on Mozambique in Box 2 for more details.

FIGURE 1. **FOOD PROCUREMENT MODELS**



Source: FAO and WFP, 2018

BOX 3. DECENTRALIZED VERSUS CENTRALIZED FOOD PROCUREMENT: SOME LESSONS FROM MOZAMBIQUE

School feeding in Mozambique was launched in 1977. From 2008, during a transition phase supported by the Brazilian Cooperation Agency and the World Food Programme, the country started promoting a school feeding programme (*Projecto de Alimentação Escolar* or PRONAE) characterized by decentralization, local procurement with direct purchase from smallholder farmer organizations, and dietary diversification in the school menu. PRONAE has been testing a decentralized procurement model at the district level. According to this model, district-level Services of Education, Youth and Technology receive the funds from central authorities and are responsible for implementing the programme locally. All food products used in the school menus (including fresh and non-perishable products) should be procured at the local level from smallholder producer associations or small traders. However, during the transition period, more centralized procurement models were implemented through the Purchase from Africans for Africa (PAA) and the Purchase for Progress initiatives.

Despite the advantages of decentralization in terms of local development, the implementation of fully decentralized food procurement is not easy, as it requires the development of human and institutional capacities at the grassroots local level. Moreover, a fully decentralized model, such as the one adopted by PRONAE, may not necessarily be the most appropriate for all types of school food products. A combination of more centralized models, such as those proposed by the PAA initiative for the procurement of certain types of products, e.g. grains, with one that is more decentralized (at the district or school level) for the procurement of fresh products have been considered as an optimal option. Such a "mixed" model could provide economies of scale and better procedural and quality control systems for the procurement of grains from smallholder farmer organizations, while combining the advantages of a more decentralized system (shorter distances and periods of storage, and increases in the variety of food) for the procurement of fresh products. For further details about implementation of decentralized food procurement programmes in Mozambique see: **http://www.fao.org/3/a-i7793e.pdf.**

Source: Authors' own elaboration.

D. INSTITUTIONAL ARRANGEMENTS

According to a 2012 survey, the responsibility for school feeding management relies on MOEs for 86 percent of the surveyed countries (WFP, 2013). There are exceptions to this general trend: in Botswana, for instance, the responsibility for school feeding falls on the Ministry for Local Government, while in Ghana it relies on the Ministry of Gender, Children and Social Protection. Alternatively, ad hoc institutions may mange the programme, as in Brazil, Cabo Verde and Chile (Drake *et al.*, 2016).

Also, there are vast cross-country variations in the interaction modalities between the institution in charge of overall programme management and other, more decentralized, functions. The lead central-level agency is usually in charge of policy formulation, standards, resource mobilization and overall management. Also targeting and monitoring are often functions of the central agency, sometimes in collaboration with local implementers (e.g. in Mexico). For a detailed overview of the core functions of

lead SFP institutions, please refer to Drake *et al.* (2016) (pp. 33-35). Importantly, there is no univocal relationship, between a given institutional setup and the procurement mechanisms adopted, to supply food. Countries present similar procurement arrangements but different ministries or agencies supervise the overall functioning of the programme.

Local authorities usually perform implementation tasks. The Plurinational State of Bolivia, Colombia and Guatemala all present decentralized management with mechanisms that vary according to their administrative policy, while Paraguay has centralized management for schools in the capital and decentralized management for the rest of the country. Depending on the decentralization level, funds are transferred either to municipalities or directly to school boards (see **http://www.fao.org/3/a-i3413e.pdf** for more details).

Food procurement models considered

To provide practical guidelines to those implementing impact evaluations, we focus on two of the most common food procurement models, being aware that other options may be practiced.

The first is a **decentralized procurement** scheme in which the food supply chain (from smallholders to schools) is contained within a small area in proximity to the procurement entity. In this case, each school procures food and vegetables from farmers living in its school catchment area. However, the smallholders who supply food usually make up a small fraction compared to the population of farmers living in the area.

In the second **more centralized model** the scale is larger. A whole district (or region) is covered by the HGSF programme and all local children are eligible for receiving the school meal transfer. In this model, food procurement occurs centrally at the district level. Similarly to the decentralized scheme, a small subset of households, whose children receive school feeding, also benefits from public food procurement as smallholder farmers supplying food to the programme.

BOX 4. COOPERATIVES OR FARMERS' ASSOCIATIONS

Procurement from cooperatives and farmers' associations, or other forms of collective contracts are common institutional arrangements encouraged to link school feeding programmes with local farmers (Bundy, *et al.*, 2009). Cooperatives can help overcome traditional barriers to market engagement for small-scale farmers by pooling together the small quantities produced by each smallholder, by reducing overall transaction costs, and by raising bargaining power to set prices and quantities. Smallholders usually provide various services, such as distributing agricultural inputs, collecting and marketing agricultural produce, conducting grading and quality control and, at times, providing transportation (WFP, 2017). However, cooperative membership may require farmers to pay fees or to conform to criteria for quality standards, which may result in less well-off smallholders being excluded from joining cooperatives.

Source: Authors' own elaboration.

E. IDENTIFY BENEFICIARIES, TREATMENT AND CONTROL GROUPS FOR IMPACT EVALUATION

HGSF are complex interventions that simultaneously reach different target population groups along the school food value chain. These may be children who receive school feeding and their respective households (especially in the case of take-home rations); smallholder farmers involved in food production; and, eventually, other community actors involved in food produce aggregation, transport, quality control, preparation and income-generating activities associated with school food provision (Gelli *et al.*, 2015).

Different food procurement modalities have different implications for the definition of treatment and comparison groups. Regarding meal distribution, children enrolled in schools targeted by the programme correspond to the treatment group in both procurement options.[6] In the decentralized model, the control group is represented by children enrolled in other schools in the same district that have not been selected for the programme. In the centralized option, since the whole district is placed under the programme, control children need to be selected from schools outside the district (e.g. neighbouring districts). Alternatively, when multiple measures of the outcomes of child schooling are available over time, the performance of children living in the same district, before and after the intervention, may be used as a control in a before-and-after analysis framework. However, this rarely occurs in practice.

BOX 5. IMPACT EVALUATION: GLOSSARY

Impact evaluation. An impact evaluation assesses the causal links between an intervention and a set of outcomes of interest.

Target population. Sectors of the population that a programme aims to reach in order to address their needs. The target population is expected to benefit from the programme. Note that the target population differs from the actual programme beneficiaries, who are those that take part in (choose to uptake or adopt) the intervention.

Treatment group. Also known as the treated group or the intervention group. The treatment group is the group of units that receives an intervention, versus the comparison group, which does not.

Control group. Also known as comparison group. A valid comparison group will have the same characteristics, on average, as the group of programme beneficiaries (treatment group), except for the fact that the units in the comparison group do not benefit from the programme. Comparison groups are used to estimate the counterfactual.

Source: Gertler *et al.*, 2016

[6] Eligibility criteria set by the programme, i.e. poverty threshold or belonging to specific population groups may eventually exclude some children from the beneficiary group, but we do not explicitly address this point here.

In the decentralized model, beneficiaries of public food procurement correspond to smallholder farmers that can act as food suppliers in school catchment areas. Therefore, treatment and control groups are formed by local farmers around treated and control schools respectively. In the centralized model, all local farmers in the district are potential beneficiaries of HGSF. However, eligibility criteria established by the programme, e.g. quality standards, minimum levels of production, or cooperative membership, may restrict the beneficiary group to population subgroups. In this case, the control group can be selected from among farmers in other districts or by applying econometric techniques that allow for dealing with selection bias of farmers in the same district who do not enrol in public food procurement.[7]

This distinction between decentralized and centralized procurement is useful for the purposes of these guidelines, as it allows for indicating how distinct food procurement models may influence the evaluation design. Thus, in the rest of the note, we will bear this distinction in mind, as described earlier on (see Section D). However, the actual differences between these two procurement models can be much blurrier in practice. For instance, school catchment areas may cover the whole district in remote areas, or schools may decide to buy a few products from farmers outside their catchment areas, because that specific food is not available locally (see case study of Mozambique in Box 3). In similar cases, the research design should be adapted to the local realities on the ground related to HGSF implementation.

As an additional layer of complexity, if farmers were organized into cooperatives or farmer associations before the intervention, this institutional arrangement may prevent them from selling their products freely on the market without involving cooperatives or farmer associations. In this case, the intervention would target these cooperatives/farmers associations as main school feeding suppliers. However, the evaluation design needs to consider that only farmers who are members of treated cooperatives would benefit from the programme. It is often the case that preferential schemes exist where procurement targets a producer organization, such as those carried out to purchase for the Purchase for Progress (P4P) programme. The implications of this possibility, related to the definition of the control group, are discussed in the following sections.

7 For a discussion of selection bias, see Section 2.2.2.

Home-grown school feeding (HGSF) impact evaluations: A stepwise approach

A five-step approach follows to assist the design and implementation of impact evaluation for HGSF programmes. The guidelines offer a set of best practices for generating evidence on the impacts of HGSF programmes on food security, farm production and schooling by taking into consideration the peculiarities of the two procurement systems presented before, i.e. decentralized versus centralized approaches.

The five-step approach is articulated as follows:

- Step 1 – *Setting up the theory of change for HGSF programmes*: identifying the channels through which the purchase of school meals from local farmers can increase agricultural profits, and in turn farmers' incomes and food security.
- Step 2 – *Choosing the research design*: presenting the main methodological tools for designing impact evaluations and the potential challenges related to estimating the actual impact of the programme.
- Step 3 – *Adopting the most adequate sampling strategies*: general recommendations for drawing a sample for impact evaluations.
- Step 4 – *Selecting outcome indicators* on different dimensions, with a focus on agriculture.
- Step 5 – *Considering implications for external validity*: discussion on whether the impact of the intervention can be extrapolated to various contexts.

STEP 1. SETTING UP THE THEORY OF CHANGE FOR HGSF PROGRAMMES

1.1 A causal chain model informed by mid-level theory

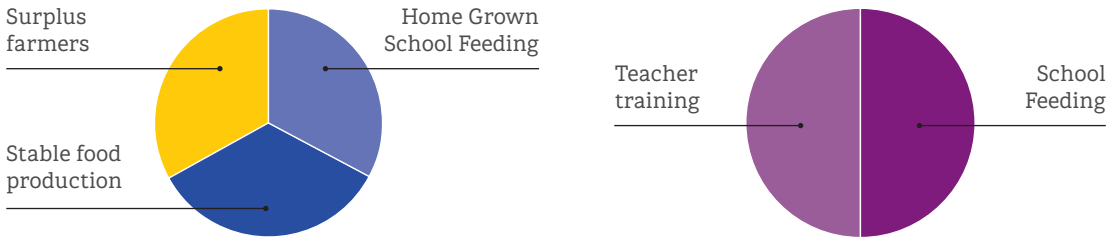
An impact evaluation design and a set of recommendations for the evaluation of development programmes should start with a good theory of change of the intervention. We use here the term "theory of change" as in Weiss (Blamey and Mackenzie, 2007) to represent both an "implementation theory" and a "programme theory". The implementation theory is the diagram representing the links between the project activities and the anticipated outcomes of the same activities. For example, the purchase of school meals from local farmers should increase agricultural revenues, which in turn should increase farmers' incomes. The programme theory comprises the mechanisms represented by the arrows linking the elements in the causal chain. For example, the demand for local food from the school meal programme may meet an elastic supply by subsistence farmers, whose resources (land and labour) are relatively unemployed and therefore readily invested in the production process. We briefly discuss three key elements of a theory of change and we then provide an application to HGSF programmes.

a. Supporting factors

HGSF interventions do not operate in a vacuum and their effects are sorted in conjunction with other operating factors. For example, an HGSF programme may increase school enrolment if, at the same time local weather conditions permit a constant flow of agricultural produce through the school year to schools, thus ensuring quality school feeding. In addition, the same outcome (e.g. increased school enrolment) can be obtained through alternative policies. For example, school enrolment can increase through many other interventions such as cash transfers, school building, and teacher training, to name a few. That an outcome can be achieved through different interventions is rather obvious, but perhaps it less obvious that the same applies to the realization of intermediate outcomes along the causal chain.

The consideration of factors supporting an intervention and alternative pathways to the achievement of the same outcomes implies an acknowledgement of multi-causality versus single causality. Multi-causality occurs in two ways. First, causes do not operate in isolation, and each cause can be seen as a set of factors that operate jointly to produce a particular effect. This is often represented as a "causal pie" as in the example in Figure 2 (Rothman and Greenland, 2005). Second, the same effect can be produced by different sets of causes. That is, each outcome can be produced with different "causal pies". For example, school feeding is a required element in a package of factors that together can increase school enrolment: school feeding is, therefore, insufficient but necessary to increase enrolment. The package itself however is only one of several packages that can produce the same outcome: the package including school feeding is sufficient to produce the result, but not necessary, as the same result can be achieved by other packages.

FIGURE 2. **TWO CAUSAL MECHANISMS OF SCHOOL ATTENDANCE**



Note: the charts present two causal mechanisms that increase school attendance. The mechanism on the left is a combination of HGSF, presence of surplus farmers in the community and stable food production in the area. The three conditions are all necessary within the causal mechanism to increase school attendance. The overall causal mechanism however is sufficient but not necessary to increase school attendance. The pie on the right is another causal mechanism (a combination of straight school feeding and teacher training) that can similarly improve school attendance and that does not include HGSF.

Source: Rothman and Greenland, 2005.

Causal chain analyses should first explore all the causes affecting the outcomes. This is true not only for the intervention as a whole, but also for each link in the causal chain of a project. If the mechanism considered is sufficiently understood, and if the supporting factors can be observed and measured, the researcher can model the negative or positive impact of these factors as "mediators" within a structural model. Second, causal chain analysis should consider how the elements of packages of interventions interact (for example, school feeding and sourcing food from local farmers), and how different activities interact with factors beyond the control of the project designer in determination of the outcomes of the intervention.

Also, timing of the evaluation matters. For instance, when an evaluation is carried out *ex ante* the delivery of the intervention, such as in an RCT or a prospective quasi-experimental study, the researchers have relative control over some of the factors affecting the outcomes, which can be manipulated in order to test hypotheses about causal relations. When an evaluation is conducted *ex post* (e.g. after the intervention has started), it is more difficult to identify the causality of specific factors, whose impact will compete with the operation of other factors that can affect the same outcomes.

At the same time, in mixed methods, when qualitative analysis precedes quantitative research, it can inform the survey design, highlighting particular themes and processes that would benefit from further examination. Alternatively, the qualitative analysis can provide immense insightful explanation to quantitative findings, by triangulating, validating and sharpening analysis of results, contributing to deeper understanding and capturing the broader experience of stakeholders. In other words, an *ex ante* evaluation has an easier task of setting out a causal chain of the intervention, because many of the alternative pathways to achieving the same outcomes can be ignored by evaluation design. For example, in designing an RCT of an SFP, we may ignore whether other interventions that increase school

enrolment are being implemented in the area. However, when evaluating a project *ex post*, the operation of other interventions, in addition to the one evaluated, need to be spelled out in the causal chain analysis. It matters, for example, whether the government was running a teachers' training programme in the same area, or whether the state was delaying salary payments to teachers, thus increasing the chance of absenteeism. It should be considered whether other schemes that purchase crops at set prices such as food reserve agencies were present in the area before or during the intervention. In that case, HGSF could lead to selling to schools only instead of these food reserve agencies without there being a significant impact on farmers' income security or productivity. In other words, in an *ex ante* randomized experiment, the implementation of other interventions is equally distributed across project and control areas and the effect of the intervention can be singled out. However, this is not normally the case in *ex post* quasi-experimental designs and the presence of other interventions needs to be considered.

The supporting factors should be separated between those that are within the control of the experimenters and those that are not. For example, the designer of an SFP may decide whether food should be sourced locally or not, but they may not be able to decide whether the area of intervention can provide food year round. The factors within the control of the experimenter should be included in the causal chain diagram. The factors outside the control of the experimenter should be included and discussed in the diagram in the form of "assumptions."

Supporting factors such as the existence of other interventions supporting given outcomes (e.g. in the case of school enrolment, a teacher training programme, or cash transfers) need to be controlled in the statistical models used to estimate the impacts of the programme. This can be achieved, for example, by modelling factors as interactions in regression analysis, or using factorial design in randomized experiments. Suppose there are two interventions (x_1 and x_2) and multiple outcomes y_i, as in equation (1).[8] In the absence of multicausality, b_3 is zero. This means that each project has an independent effect on the outcome and the effects are additive. For example, school feeding (x_1) and teacher training can both have an impact on school enrolment, b_1 and b_2 respectively. The two effects can be totally independent in such a way that for school feeding to have an impact on enrolment it does not really matter whether teachers are trained or not, and vice versa.[9]

$$y=\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2+\epsilon \quad (1)$$

In this case, x_1 and x_2 are non-necessary but sufficient single causes of the outcomes (b_3=0). This is normally a special case and more common is the multi-causality case (b_3>0), whereby projects require other conditions to be present to have an impact. These latter projects are therefore INUS conditions and several cases are possible.[10] First, projects may have an impact only if implemented together (b_1=b_2=0 and b_3>0). For example, teacher training might have an impact on learning only if children

8 This treatment borrows from Ravallion's discussion of the impact of portfolio interventions (2015).

9 A special case occurs when both projects have an independent effect on the outcomes, but their impact are not additive and operate to the exclusion of the other. For example, school feeding has an impact on school enrolment, but the impact is zero if there is a food distribution intervention. Defining p=1 when school feeding is operating and p=0 when not, the impact of the two interventions is:

$$y=\alpha+\beta_1 P x_1+\beta_2(1-P)x_2+\epsilon$$

10 INUS conditions are "Insufficient but necessary parts of a cause that it is unnecessary but sufficient. Such are "home-grown school feeding" and "school feeding" in the pie examples of Figure 2.

are going to school, which requires an incentive programme such as conditional cash transfers. In the same way, conditional cash transfers may not have an impact on learning unless teacher training takes place. Second, project x_2 may not work independently and require the operation of project x_1. Finally, the projects may have additive effects and also interactions ($b_1>0$, $b_2>0$ and $b_3>0$).

All the above examples refer to activities that are under control and that can be altered at the evaluator's will. But the same reasoning applies to factors outside control, such as for example the parents' levels of education or quality of higher education, which may have an interaction effect on the outcomes. These factors cannot be manipulated, at least in the short term, and therefore have limited relevance to the design and evaluation of a specific intervention. But they matter for its external validity, as the activities will have the expected impact only when these other factors are present.

In summary the main recommendations for the exploration of supporting factors are the following:

- The analysis should consider alternative pathways to achieve the same outcomes whether they are intermediate or final: what are the different causal packages operating at each link of the causal chain? Note that at each different step of the causal chain, different causal mechanisms can be in operation so that several causal pathways are possible.
- The analysis should include supporting factors in the diagram that are under the control of the experimenter and (somehow) should include and discuss as "assumptions" those that are beyond control. Factors that are outside the control of the experimenter can be considered at the analysis stage, for example as control variables in a regression model, but cannot be included in the causal diagram as they cannot be affected by the intervention.
- The analysis should formulate causal packages as interactions and suggest ways of testing the interactions, for example through factorial designs, regression analysis or path models.
- In a mixed method impact evaluation, the overall design should consider sequencing of methods to identify the optimal approach to maximize the value of each method. As well, from the start of design, it is important to address how results will be integrated most effectively to explain impacts employing triangulation of methods.

b. Mechanisms

Analysis of mechanisms consists of two parts: modelling mechanisms; and designing evaluations, which are testing mechanisms. First, all the causal relations of the theory of change need to be described and explained. This step must be conducted using behavioural models showing how an activity is transformed in an outcome and under what conditions. Economic models are set up to set out causal relationships in the theory of change, but alternative models can also be employed, such as for example, mathematical modelling, path analysis and structural equation modelling. The assumptions for a behavioural mechanism to operate in the expected way should be explicit.

Second, the analysis should identify weak links in the causal chain. Weaknesses refer to causal links that are not completely understood or for which there is little evidence available. Evaluations may set out to understand and to evaluate these links rather than estimating the extent of the impact of the intervention on the final outcomes (Ludwig *et al.*, 2011). This is another form of theory-driven evaluation whose goal is to uncover behavioural responses and processes that work in similar ways in different contexts.

For the result of a study to be valid outside the original area of application (external validity), the causal mechanism needs to be understood well. The causal mechanism may contain different degrees of generality in its application and validity. Some mechanisms are so general they are always true but of limited use. Conversely, some mechanisms are highly contextualized and only local to the context considered. The latter mechanisms are very informative of how the intervention works in a particular context but do not help in understanding the applicability of the same intervention in other contexts. What is needed is a mid-range mechanism located at a level of generality that it is valid across contexts, but at the same time is not so general that it is no longer useful. The qualitative method is particularly valuable in complement to the quantitative analysis in causal analysis. In this case, greater depth of understanding, for example about perceptions, attitudes, norms and behaviours, can contribute to increasing comprehension of causal processes and pathways of impacts.

For example, in an HGSF programme, the mechanism for farmers to increase income is profit maximization. According to economic theory, farmers operate to maximize profits. Hence, if offered the right incentives, farmers will increase food production. This causal mechanism appears to be too general. Risk-averse farmers may maximize expected utility and hence prefer a less variable income stream to a more profitable but more variable production activity. The causal mechanism is valid if we do not bother describing the incentives being offered by the programme. On the other hand, we could set out to describe the theory of change in the Ghanaian HGSF project (Gelli *et al.*, 2016). In Ghana, farmers would not respond to incentives. The reason was that in Ghana caterers provided the school meals and were paid after the meals had been served. Caterers were unable to purchase food from farmers in advance and instead had to rely on food credit from "market queens". Farmers could have sold their produce to caterers on credit, but the level of trust between farmers and caterers was not sufficient to make this type of arrangement possible.

The lack of credit mechanisms, enforceable contracts, lack of mutual trust, led to the absence of a market in which transactions between farmers and caterers could take place. While the programme may work in principle, it does not in the absence of some necessary supporting factors. The theory of change in the Ghanaian project, therefore, describes an intervention that facilitates contractual arrangements between caterers and farmers. The same arrangement, however, may not work in other contexts where there are no "market queens," or where caterers do not purchase the food. Knowing that the Ghanaian model worked therefore does not guarantee that it will work elsewhere, although lessons can be learned that can be adapted to other contexts. What is required is a mid-range theory of farmers' participation in the programme. Farmers will provide food to the programme, but only if there is a market where the transaction can take place. Spelling out the causal mechanism helps us understand the circumstances in which the programme is expected to work.

Identifying the mid-level mechanism implies identifying mechanisms that "work" under a range of different contextual conditions. Once the general causal mechanism is defined, the goal of the project designers is to adapt the interventions to the specific characteristics of each context. A specific project will not work in the same way everywhere, but variations in a valid causal mechanism can be researched and identified.

In summary, the main recommendations in relation to mechanisms are the following:

- Each link in the theory of change needs to be explained using a behavioural model and the assumption arrived at should be explicit.
- Weak links in the theory of change need to be identified and suggestions made about how they should be understood or tested, including through a mixed methods approach.
- The theory of change should be formulated at a level of abstraction sufficiently general to be valid across contexts with proper adaptations.

c. Context

We define the context as the set of covariates describing a population, as well as the complex systems of norms, institutions and relations that support a causal pathway. When the characteristics of the context are well known and understood, effects of interventions can be extrapolated exploiting the heterogeneity of characteristics of population, interventions, and locality using statistical methods. The simplest type of context is the one in which the factors supporting the interventions are known, can be observed and measured. In this case extrapolation of programme effects over distributions of contextual factors, such as for example education levels of beneficiaries, are possible. Other types of contexts however encompass social relationships, power relations, or norms, which are more difficult to observe and measure.

We therefore propose a definition of context that has an increasing level of complexity, from simple, to difficult, to complex. Simple extrapolation of the covariates and subgroup analysis may be sufficient in simple contexts, while more complex contexts may require an analysis of central features of the contexts or of markers. For example, extrapolation of the effects of HGSF programmes may require knowledge of the market structure or other key characteristic of the context, which allows the incentives to operate. Highly complex programmes, for example national health services or "foreign aid" as a whole, may be approached through markers, such as for example the claim that foreign aid works in the presence of "good governance," where good governance becomes a marker of success for exporting a successful intervention from one context to another.

A related element in the context, which is relevant to any analysis of the external validity of an intervention, is an examination of its scalability. Some interventions are characterized by effects that vary with the scale of operation. For example, a vaccination programme is more effective when conducted at a large scale, while a cash transfer programme may face increasing costs as the scale of operation increases. In these cases, the researchers need to assess whether a partial equilibrium analysis is sufficient to detect project effects. Note that all impact evaluations are conducted in partial equilibrium and effects cannot always be extrapolated to a larger or national scale. Researchers need to conclude whether a general equilibrium analysis of the intervention is needed.

In summary, the main recommendations from an analysis of the context are the following:

- Characterize the context as simple, difficult and complex depending on the level of complexity and knowledge of the underlying structural mechanism.
- Depending on the classification made adopt:
  - extrapolation of effects on the relevant covariates at population, project and locality level (subgroup analysis);

- identification and discussion of the central features of the structural mechanism;
- identification of markers.

- Assess the scalability of the intervention and recommend whether a general equilibrium analysis, or other type of analysis, is needed.

1.2 A theory of change for HGSF programmes

Supporting factors

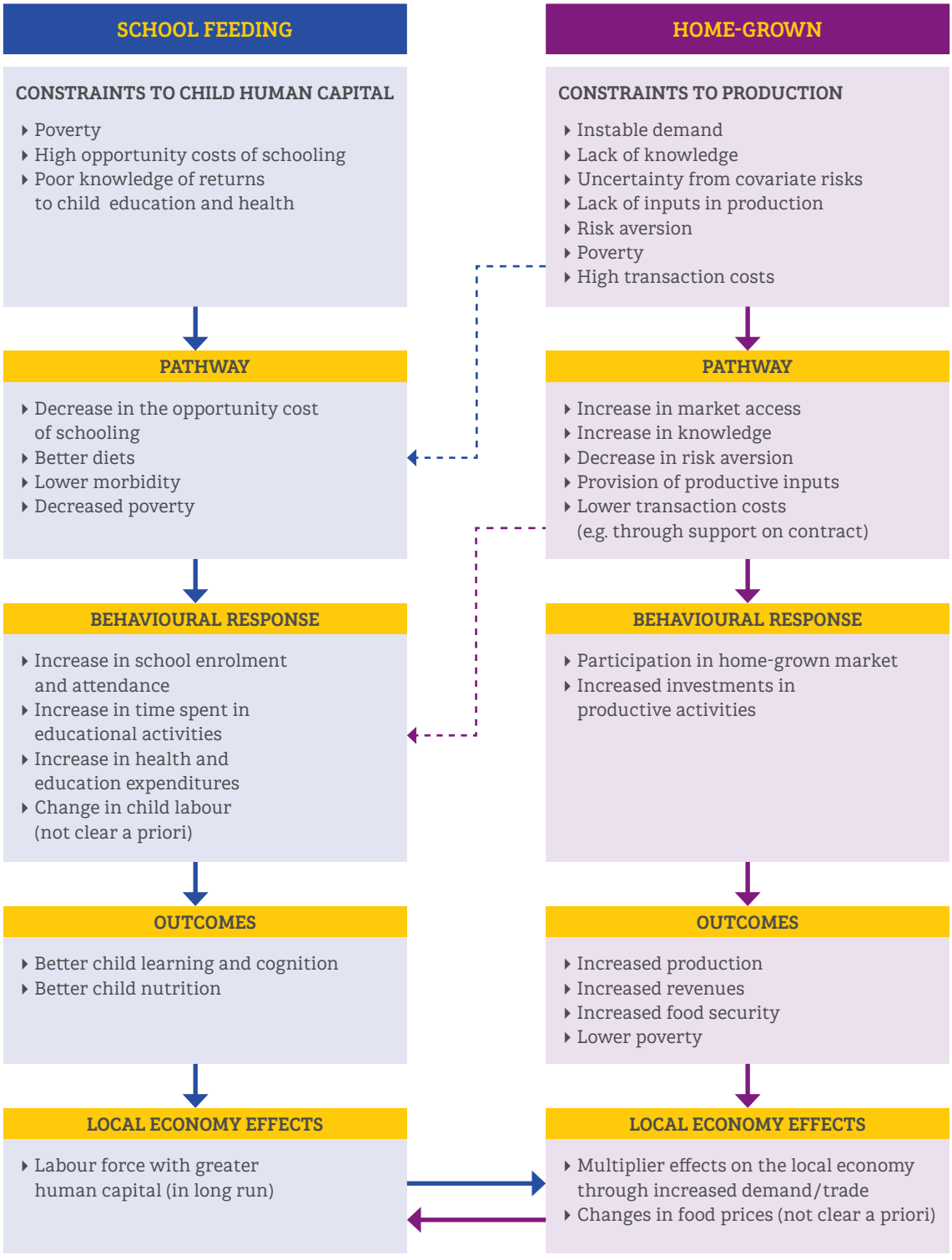
An intervention does not occur in isolation and, while some of the supporting factors are under the control of the policymaker, other factors are not. In this section we focus on the factors that are under the control of the policymaker and we will discuss factors outside control in the section on context.

Figure 3 is a highly stylized high-level theory of change in HGSF interventions. The chart, as well as much of the discussion in this section, draws on FAO and WFP (2018) as well as on previous studies. The programme aims to provide benefits to school-going children, farmers and the wider local community. The benefits for children are those provided by standard SFP. These benefits are known and proved (see Section B), and are divided into education benefits (higher school enrolment, attendance, cognition and learning) and health benefits (improved nutritional status and general health). Benefits to farmers consist of increased incomes, food security and changes to agricultural production patterns. The home-grown component of the programme generates an additional demand for food, which increases farmers' incomes or decreases income variance over time (better income security). To the extent demand is stable over time, it may also help stabilize income against seasonal fluctuations and other shocks, promote food security, and encourage agricultural investments.

Sometimes, other more indirect benefits are mentioned such as increased access to markets, increased access to credit, increased institutional capacity, reduced post-harvest losses, and benefits from joining cooperatives. In the long term, higher income and improved expectations regarding future income streams may also increase farmers' investments in their farms or in their, or their children's, human capital. The second stream of benefits to farmers consists of diversification of production towards foods that are healthier and more nutritious for the school food supply. Many of these benefits are only hypothetical. An increase in income through an expansion of demand may be the only truly likely effect of HGSF programmes, while higher income stability, farm investments, and production diversification will occur only under a few conditions. For example, it has been observed in Zambia that beneficiary farmers would reallocate resources within the existing production mix, rather than diversifying into new crops, presumably because of land, labour, inputs and other constraints that limit their capacity to meet the extra demand without reducing parts of other productive activities. Finally, the wider community may benefit, particularly traders, caterers and food processors through increased income and job opportunities.

Note that the arrows in Figure 1.2 link nutrition and education, as well as production of nutritious foods and child nutrition. There is evidence that better-fed children are more likely to attend school because they miss fewer days due to sickness, and learn more while in school because they are more attentive, which explains the arrow running from child nutrition to child education. It is also believed that children are more likely to accept a more diversified school meal, based on nutritious local food. If, local production is diversified it will be more effective from the viewpoint of nutrition.

FIGURE 3. **HIGH LEVEL THEORY OF CHANGE IN HGSF INTERVENTIONS**



Source: Authors' own elaboration.

If we call the benefits to children as "school feeding" intervention and the benefits to farmers as its "home-grown" component, we can proceed to consider how the effects of the two components interact. We do this exercise for three outcomes: school attendance and learning; nutritional status; and incomes, which are summarized in Table 1. Both school feeding and home-grown components directly impact schooling. The impact of the school meal is obvious, more children will go to school if offered free meals, thus increasing school enrolment and attendance, and potentially raising learning and cognition through better schooling and nutritional status. The impact of the home-grown component on schooling is more indirect and occurs through an increase in the family income of those farmers supported by the programme. Part of this income will be invested in child education and nutrition following household spending decisions.

The size of the impact of this component will depend on the number of farmers supported at the local level, thus on the income effects on local farmers participating in HGSF, as well as on the farmers' prioritization of child education and health in face of increases in their disposable income. The interaction effect, resulting from jointly implementing school feeding and its home-grown component, occurs through diversification of diet. Locally produced food may be healthier and more diverse, which may positively impact child nutrition and school attendance. However, a negative interaction effect may be possible if the HGSF programme increases the use of child labour on the farm to supply additional food to the schools, thus decreasing school attendance or time in educational activities.

The impacts on nutritional status are similar. School meals improve nutritional status directly as children consume more food and, in the home-grown component, children consume more food as a child's household becomes better-off because of the programme, thus potentially increasing its expenditure on food consumption and on more diverse diets. The joint provision of school meals and of farmers' incentives (the interaction) acts on nutrition through an improvement in the diet, if locally-produced foods are sufficiently diverse.

The impact of the two programme components on household income is different from the case of child education and health. School feeding has a direct positive impact on incomes to the extent that families can reduce expenditure on children meals, which is the social protection aspect of school feeding. Like other social protection interventions, school-feeding acts as a fungible (although indirect) transfer equivalent to the size of the meal consumed at school that households can spend in different ways. The impact of the home-grown component on farmers' incomes is obvious: larger production and profits increase farmers' incomes. But the two components do not produce a positive interaction effect.

TABLE 1. **IMPACT INTERACTIONS OF SCHOOL FEEDING AND HOME-GROWN COMPONENT BY OUTCOME**

	School feeding	Home-grown	School feeding*Home-grown
School attendance	(+) direct	(+) via income	(+) via change in diet (-) via child labour
Nutrition	(+) direct	(+) via income	(+) via change in diet
Income	(+) direct	(+) direct	Not applicable

Source: Authors' own elaboration.

Notice that the income effect of the home-grown component can be assessed without the school-feeding component. If the study is interested in evaluating the impact of the home-grown component on farmers' incomes, it can be done independently of the school feeding intervention, thus simplifying the overall evaluation design. This is because there is no expected impact on income of the joint implementation of the home-grown and school feeding components. This reasoning is valid in the short run only, because in the long run a healthier and better-educated work force (stemming from the school feeding component on child human capital) will result in higher incomes and productivity – but this cannot be detected within the short timeframe of a typical impact evaluation.

On the other hand, if education and the nutritional status of children are of interest, then the evaluation of school feeding, home-grown school feeding and their interactions should be carried out simultaneously in order to disentangle the direct effects of the two components and the interaction effects. Of course implications arise for the final sample size of the study and complexity of the study design (e.g. two-level randomizations examining the effects of agriculture components at a different level from effects on education/nutrition).

It is also suggested that HGSF needs to be integrated into other sector policies to be effective (FAO and WFP, 2018). Interventions in agriculture to remove bottlenecks and constraints are considered necessary. It is also sometimes suggested that HGSF should be one element of a broader package of interventions to improve health and education. In the case of agriculture, additional intervention may occur in infrastructure (irrigation, storage, roads, electricity, processing facilities), inputs (seeds, fertilizer, machinery, transport, credit and insurance), agricultural practices (changes in farming practices), and the environment (land titling, legal framework, etc.).

HGSF interactions with other projects raise several other interesting questions along the same lines as the analysis described above. Are the hypothesized synergies likely to materialize and can they be estimated using interactions? What it the size of the interaction of HGSF with other interventions? Are they really needed for the success of HGSF? What is the independent effect of HGSF? Can HGSF be considered and evaluated as a standalone intervention without other elements in the package?

Causal mechanisms

Our approach to the exploration of causal mechanisms has the goal of generating several questions and hypotheses underpinning the assumptions behind the implementation of HGSF programmes. Answers to some of these questions are likely to be valid for "small farmers in poor rural areas," and therefore are valid beyond the specific context analysed.

The impacts of HGSF interventions on farmers' income were discussed in Masset *et al.* (2013) and will only be summarized here. We focus on three critical questions:

- Will the intervention generate additional food demand in the local market?
- Will local farmers respond to the additional demand by producing more food and increase their revenues?
- Will farmers make investments in their farm to respond to the larger and more stable demand for food induced by the intervention?

It is not obvious that the purchase of food from local farmers will result in an increase in demand for the same amount. For example, households may reduce consumption of locally produced food by the same amount as the school meal in such a way that the net effect on the market is nil. In this case the effect of the project would be to transfer demand from a few local farmers to other local farmers but without creating any additional demand. Note that the result of a full substitution effect is to turn the benefit of the project in favour of consumers rather than producers. With full substitution, consumers increase savings that they can spend on food or non-food items, while on balance producers do not gain. It is likely that the reduction in the consumption of local food occurs to some extent, but not to the point of cancelling out the increase in demand produced by the project. An understanding of this process requires knowledge of incomes of all local farmers (not just those providing food to the HGSF programme) and of consumption patterns in households.

Depending on the characteristics of local markets, and of the contractual arrangements made with local producers, the intervention may generate an increase in food prices, at least for the producers supplying the project (see Figure 3). Whether farmers will respond to price incentives will depend on several factors including: their ability to produce more food, which could be compromised by seasonality and lack of storage facilities; the ability to mobilize more inputs (labour, land and fertilizer), which may be constrained by lack of credit access or not adjustable in the short term. In other words, the size of the farmers' "supply response" is uncertain. An understanding of this process requires knowledge of existing farming technology, input availability and market prices.

Whether farmers will increase farm investments will depend on the credibility of procurement arrangements and of prevailing perceptions of risk. It is normally assumed that agricultural risk is high and that farmers are risk averse. If this is the case, investments are constrained by actual risk and farmers' attitudes to risk. However, it is difficult to predict farmers' response to a more stable income without knowledge of existing production risk and of farmers' risk aversion. An understanding of farmers' response to risk reduction requires an understanding of farmers' attitudes to risk and the prevailing risks to production in the area.

Take-away message

Answering questions about causal mechanisms determining the impact of HGSF programmes also places demands on the type of data that the evaluation should collect. Data collected should include: household incomes and expenditures data; purchase and selling prices for local foods; access to markets, credit and other productive inputs; and farmers' attitudes to risk.

The context and its implications for external validity

Before embarking on programme evaluation, the evaluating team should understand the context in which the programme operates by undertaking a qualitative analysis of agricultural markets and the institutional capacity to implement the programme at various levels. A preliminary analysis maps the key features of the intervention, e.g. food processing model, acceptability of the intervention among beneficiary groups, is also recommended before starting the evaluation.

Understanding the context in which the programme operates is critical to extrapolating the results of the study to other settings. Although HGSF programmes are not as overly complex as those that include multiple actors and provide highly differentiated services, e.g. national health service or foreign aid, they are still characterized by moderate complexity, which does not allow for the straightforward extrapolation of results. In other words, impacts observed in one context cannot be simply extrapolated to other contexts. In addition, characteristics of the farmer population, school children, communities, etc., which are normally collected through surveys, cannot be used on their own to extrapolate results from one context to another. An assessment of the ability to replicate results in different contexts requires an understanding of key contextual factors. Successful projects can be replicated in those places where the same contextual factors supporting the intervention are present.

In particular, the programme will not be effective in the absence of a supporting operating model and procurement system. A suitable procurement model linking local farmers to the programme is key to the success of the intervention. In turn, the procurement model is part of a larger operating model including all the actors engaged in trade, preparation and distribution of foodstuffs. Different operating and procurement models are possible, each being supported by different contextual factors. The starting point of an investigation of the transferability of results from one context to another (the programme "external validity," see also Step 5) is, therefore, a context analysis of the factors supporting viable operating and procurement models.

STEP 2. CHOOSING THE RESEARCH DESIGN

This section provides the reader with methodological tools for designing impact evaluations, taking into consideration the potential issues that can hamper estimation of the real impact of the programme in the context of HGSF interventions. First, we describe the general objectives and core concepts related to impact evaluations. We then present an overview of common methodological designs, i.e. experimental and quasi-experimental methods. In the second part of the section, we discuss the mixed methods approach, specifically the qualitative approach. We also address the main methodological challenges of designing evaluations of HGSF programmes, by proposing many suitable techniques that will address these challenges. Methodological guidelines are separately elaborated for decentralized and centralized procurement models.

2.1 Definition and aims of impact evaluation

The goal of an impact evaluation is to empirically verify to what extent a programme contributed to a change in selected outcomes (e.g. schooling, nutrition, food security). To establish causality between an intervention and a given outcome, impact evaluation methods need to rule out the possibility that any factors other than the programme itself explain the observed impact. For instance, in the case of SFP that focus on education, what is offered by the programme should be the *unique* possible explanation for any difference in a child's schooling outcomes. To eliminate the influence of potential

confounding factors, the evaluation should measure each outcome at the same point in time for the same unit of observation in two different states of the world, i.e. going back to the earlier example, whether the same child has been exposed and not being exposed to the SFP ("counterfactual"). This is, by definition, impossible in practice.

To gain evidence of the impact of the programme from simply comparing outcomes for programme participants before-and-after treatment, we need to assume that, in absence of the programme, outcomes for participants would have been the same as they were before the programme. Unfortunately, in most cases, this assumption does not hold. Similarly, comparing a group of individuals, who select to take part in a programme with others, also would not produce a rigorous comparison, as programme participation is based on preferences, decisions, or participants' unobserved characteristics, which, in turn, are likely to affect the ultimate outcomes of the evaluation (Gertler, *et al.*, 2016; White and Raitzer, 2017).

The group that is assigned or exposed to the programme is known as the "treatment group" (See Box 4 for definitions). To estimate the impact of the programme, it is therefore critical to retrieve a counterfactual for this group. In other words, we need to identify a group that is statistically identical in all other characteristics to the treatment group, except for exposure to the programme. The "comparison group," by remaining unaffected by the programme, allows us to estimate the counterfactual outcome, that is, the outcome that would have prevailed for the treatment group had it not taken part in the programme (Gertler *et al.*, 2016). Identifying such comparison group is a crucial step in any impact evaluation, regardless of the type of programme being evaluated.

To identify a comparison group, the analyst needs to understand the mechanism underpinning the assignment to the programme, i.e. the process by which those who receive the intervention are selected and can self-select to participate in the programme. The key challenge in the identification of a suitable comparison group consists of the fact that usually the people who are selected or choose to participate in the programme are usually not the same, in terms of observable and unobservable characteristics, as those who do not. They may be better informed or educated, more willing to take risk, more proactive, or have other behavioural differences from those who do not participate. For instance, in the case of HGSF, farmers who are offered the chance to participate or self-select, may have larger fields, greater access to credit, and better production capacity than farmers who may not be offered the chance to participate or chose not to.

Take-away message

In order to guarantee an ideal counterfactual, treatment and comparison groups should fulfil the following properties the:

- average characteristics of the two groups must be identical in the absence of the programme;
- treatment should not affect the comparison group either directly or indirectly;
- outcomes of units in the control group should change the same way as outcomes in the treatment group, if both groups were offered the programme (or not).

A control group that differs from the treated group in other ways, other than the absence of the treatment, would generate biased estimates as to the effect of the programme that would be mixed with these other differences.

An impact evaluation may provide different estimates of intervention effects, depending on the sample from which the estimate is generated. This is important to consider when designing an evaluation, as not all methodologies can estimate all measures. Box 6 summarizes the definition of different impact measures.

BOX 6. DEFINITION OF IMPACT MEASURES

- **Average treatment effect** (ATE): average impact of participation in the programme on the entire eligible population.
- **Average treatment effect on the treated** (ATT): average impact on those who actually take part in (choose to uptake or adopt) the intervention.
- **Average treatment effect on the untreated** (ATU): the average potential impact on those not taking part in the treatment were they treated. Relevant to understanding the potential effects of programme expansion.
- **Local average treatment effect** (LATE): average impact on a subgroup of the beneficiary population, usually those at the threshold of eligibility.

When all units eligible for treatment take part in the programme, ATE corresponds to ATT. For instance, if all children in a given school are offered school feeding and take up the intervention, we would have equivalence between ATE and ATT. Thus, eligible and treated populations overlap. However, in most cases, where participation in the programme is voluntary (also called "imperfect compliance"), ATT and ATE diverge. For instance, although they are offered the programme, children may enrol in another school or drop out, thus leaving only a subset of children (the "treated") receiving the programme. Similarly, farmers in school catchment areas selected for home-grown programmes can decide whether to sell their products to schools or not. Hence, as typically those who self-select into interventions gain more than those who do not, the treatment effect on those farmers who decide to take part in the programme would be larger than the ATE estimated on the entire eligible population of farmers in the area.

In general, the following relationship should be observed among impact indexes: ATT>ATE. Indeed, the average effect on the target population will be necessarily no greater than the average effect on those who actually take part. An intervention may have a very large impact on those actually taking part, but if only few of the people who are offered the programme actually take it up the ATE effect may be low. Nonetheless, estimating ATE is still highly relevant to policy, as most times not all units offered an intervention will accept it, thus ATE provides a measure of the effect of the programme in presence of imperfect compliance (for further discussion on imperfect compliance, see Box 10).

Source: Authors' own elaboration.

2.2 Evaluation tools: how to design a counterfactual?

Several empirical methods can be applied to construct the counterfactual. The choice of the preferred strategy depends on multiple factors, such as context, available data, or the possibility of intervening in the design of the intervention from its early stages. Both experimental and non-experimental designs are suitable for impact evaluations of HGSF programmes. Section 2.2.1 describes a range of *experimental*

methods. The randomized assignment of treatment in experimental methods will ensure the internal validity of the results, as the comparison group provides an accurate estimate of the counterfactual that allows estimation of the true impact of the programme. External validity requires that the evaluation sample accurately represents the population of eligible units (see Section 3 on sampling strategies). In this case the results of the evaluation can be generalized to the population of eligible units (Gertler *et al.*, 2016).

2.2.1 Experimental designs

Randomized experiments are commonly considered as the "gold standard" method for conducting impact evaluations. An RCT involves the random assignment of members of the eligible population to one or more "treatment groups" that receive the intervention, and to the "control group" that receives no intervention (see Figure 5 in Appendix A). When the number of potential beneficiaries is larger than the number of participants that the programme can serve, the randomized assignment often comes directly from the programme's operational rules. For instance, when resources are sufficient to provide school meals to all children enrolled in primary schools in a given area, randomization can determine which of the eligible pupils are included in the programme, and which of them should be assigned to the control group.

In the specific case of school feeding, this could be achieved, in practice, by randomly assigning school feeding to a subsample of eligible schools in the area, leaving the rest as a control. Alternatively, if a programme needs to be gradually phased in until it covers the entire eligible population, the control group can be constructed by randomizing the time in, which participants are enrolled in the programme. This randomization scheme is well known as a "pipeline" or "step-wedged" design. As long as the last group has not yet been phased into the programme, it serves as a valid comparison group for those who have already been phased in. This setup, called stepped-wedge trial, can also allow estimation of the effects of differential dose exposures to treatment.

The random assignment of units to treatment and control groups has a high probability of generating two (or more depending on the number of treatment arms) statistically identical groups. In general, if the population of eligible units is large enough, the randomized assignment mechanism will transfer any characteristic of the population to both the treatment and the comparison group. This is expected to be the case for both observed characteristics, e.g. individual or household socio-demographics, and unobserved variables, such as motivation, preferences, or other features that are more difficult to measure. Thus, treatment and comparison groups that are generated through randomized assignment will be similar not only in their observed characteristics but also in their unobserved characteristics (Gertler *et al.*, 2016).

Baseline data collected before the intervention rollout can be used to verify this assumption empirically, by testing for no systematic differences in observed characteristics, and the study outcomes between treatment and comparison groups. If this condition is satisfied, and the two groups are exposed to the same external environmental factors over time, differences in outcomes between the treatment and comparison groups after the launch of the intervention can be ascribed to the introduction of the programme. The comparison group controls for all factors that might also explain the outcome of interest. **Under the randomized assignment, the impact of the programme is simply given by the difference between the mean outcome of the treatment group and the mean outcome of the comparison group**.

This estimation is unbiased, since all observed and unobserved factors that might otherwise explain the difference in outcomes are ruled out by randomization (Gertler *et al.*, 2016).

The random assignment can occur at different levels. In more simple contexts, the unit of assignment is the same as the unit of treatment and measurement. In some contexts, to reduce the risks of spillover effects,[11] imperfect compliance or contamination, a cluster RCT design may be preferred. In this case, the unit of randomization contains multiple treatment units. For instance, randomization can be conducted at school or village level, and all individuals or households in that unit are assigned to either treatment or control group. Knowledge of treatment, however, could spread within and outside a community, altering the behaviour of the untreated and generating estimation biases. Cluster RCTs can cope with this by creating large and distinct enough units of assignment so that spillover is minimized. The statistical power of the design depends on the number of clusters in the study rather than the number of treated units. This means that the example programme will have to cover a reasonably large number of treatment villages, communities or schools to obtain a sufficient power (White and Raitzer, 2017) (see also discussion on power analysis in Step 3).

Practical tips!

See general guidelines to conduct a randomized control trial see:

- White and Raitzer (2017), Chapter 3
- Gertler *et al.* (2016), Chapter 4

BOX 7. IMPLEMENTING EXPERIMENTAL DESIGNS: DATA ANALYSIS METHODS

When the design of an RCT is rigorously conducted and the evaluation sample is randomly representative of the population of interest (see Step 3), estimating the impact of the programme is relatively straightforward. After a period of programme implementation, outcomes for both the treatment and comparison units need to be measured. The impact of the programme is simply the difference between the average outcome for the treatment group and the average outcome for the comparison group, which can be calculated by using multivariate regression models. Instrumental variable methods using random assignment as an instrument can be used to deal with imperfect compliance.

For an overview of data analysis techniques for RCT, see:

- Duflo, E., Glennerster, R. and Kremer, M. 2007. Using randomization in development economics research: A toolkit. *Handbook of development economics*, *4*, 3895-3962.
- Read, K. L., Kendall, P. C., Carper, M. M. and Rausch, J. R. 2013. Statistical methods for use in the analysis of randomized clinical trials utilizing a pretreatment, posttreatment, follow-up (PPF) paradigm. *The Oxford handbook of research strategies for clinical psychology*, 253-260.

Source: Authors' own elaboration.

[11] For further details, see the definition of spillover effects and imperfect compliance in Box 10.

2.2.2 Quasi-experimental designs

Despite the advantages in terms of retrieving a plausible estimate of causal effects, conducting randomized experiments can be problematic for many different reasons. Implementing agencies may not be willing to pursue randomization for ethical or organizational issues, or the type of intervention does not allow for randomization or sequencing. Further, the programme may have already started when evaluation is designed, or there may not be the funds to conduct such evaluations, as RCTs are usually quite costly. When randomization is not possible, a range of *quasi-experimental designs* can be adopted.

These methods rely on forming a comparison group (thus, a valid counterfactual) by employing statistical methods, rather than exploiting the random assignment to treatment and control groups. **If treatment and control groups are rigorously constructed, the average values of observable characteristics should be equivalent, as happens in RCTs, and thus estimates of treatment effects are unbiased. The main methodological challenge to adopting these methods is represented by unobserved characteristics of the unit of analysis that may simultaneously influence participation in the programme and the outcomes, which is usually referred to as "selection bias."** Several types of procedures can be adopted to design evaluations that are able to deal with selection bias.

One of the most robust quasi-experimental methods is the **Regression Discontinuity Design** (RDD). This design is often used when programmes entail a threshold to establish eligibility for assignment to the intervention. The threshold is usually based on a continuous assignment variable, which is measured for all potentially eligible units, i.e. a poverty index, household income, school performance. This method builds on the idea that units immediately below the threshold are sufficiently similar to those above the threshold, and thus included in the programme, to be a valid comparison group. The difference in outcomes between those near either side of the threshold corresponds to the measure of impact. RDD is considered a good method for ruling out the threats to internal validity posed by unobservable factors that simultaneously influence both programme enrolment and outcomes. Importantly, beneficiaries should not be able to manipulate the assignment variable to become eligible for the programme as such manipulation would generate selection bias.

In order to obtain unbiased estimates of the treatment effect, the relationship between the assignment variable and the outcomes should be parallel in the treatment and control group, and researcher's assumptions about the functional form of this relationship need to be met. The technique requires the evaluator to have a clear idea of the assignment criteria and sufficient sample size at either sides of the threshold. In addition, the threshold should be unique to the programme evaluated otherwise we cannot disentangle the effect of the specific intervention. A key limitation of RDD is that the impact is estimated only for the population close to the threshold, therefore it provides a measure of the local average treatment effect (LATE) rather than the ATE, with remarkable implications for external validity (Campbell and Stanley, 1963).

Interrupted Time Series (ITS) is an alternative application of RDD in which the threshold is the point in time at which a programme comes into effect. ITS are typically used to evaluate the effects of an institutional change or a social innovation, where the intervention impact is expected to be sudden, rather than gradual (e.g. a change in the length of the school year, which is enacted for some birth cohorts, but not for others). In order for this design to be applied, data on the outcome variables

need to be measured at equally spaced intervals, e.g. daily or yearly, over a long period. In this way, analysts can identify the changes in the level and slope of the series occurring in correspondence with intervention rollout. In other words, the expected value of the treatment group can be compared with the expected value of the control group, conditioned on the point in time at which the treatment is introduced (Reis and Judd, 2013). To go back to our example, related to the change in length of the school year, this design implies that we would compare the educational outcomes of the birth cohorts exposed to the change as compared to the ones for which the length of the school year is unvaried.

As for RDD, the correct functional form of the relationship between time and outcome needs to be specified and parallel trends in the outcomes between treatment and control groups, in absence of treatment, need to be assumed to obtain unbiased estimations. In addition, time should be a good proxy for the actual rule upon which treatment and control conditions are assigned (Reis and Judd, 2013). Common threats to internal validity in this setting are: other events that are beyond the researchers' control that occur at the same time as the programme and may affect the outcome; selection issues related to people knowing about the intervention before and making an effort to be enrolled; measurement error issues caused by changes in the recording procedures occurring around the timing of the intervention; anticipation or delay in actual implementation of the change that is not observed by the researcher (Reis and Judd, 2013). When more than one replication of time series data are available, e.g. data from children experiencing changes in the duration of the school year in different regional contexts and modalities, the external validity of impacts estimated through ITS can be tested.

A third option is the **Non-equivalent Control Group Design**. In this design, a group of participants is either given a treatment or they experience a "natural" event, such as a change in policy, exposure to natural shock, etc. Thus, a comparison group that does not receive the treatment has to be identified by researchers. However, since the process determining whether participants end up in the treatment group or not is not fully known, it cannot be statistically modelled. A typical example of the application of this design is to compare the academic achievements of students who attended private versus public schools. As for randomized experiments, to produce unbiased estimations of treatment effects, treatment and control groups should be equivalent, on average, in terms of all relevant background characteristics before the intervention (pre-test characteristics).

In addition, variation in the outcomes of interest between pre and post-tests should be the same across groups in absence of the programme. However, since treatment status is not randomly assigned, we can expect pre-test characteristics to be related to treatment assignment. Thus, researchers must use all available information to identify background characteristics so as to model the process of selection for treatment, while all other covariates and unobserved variables are assumed to be unrelated to treatment status (assumption of "strong ignorability") (Campbell and Stanley, 1963).

The latter can be quite a strong assumption to hold. Researchers should identify important covariates by considering all variables that are theoretically expected to be related to selection for treatment, as well as variables that are known from the empirical literature to be related to the outcomes of interest (for details on how to operationalize this technique, see discussion on matching technique in Box 8). Violation of the strong ignorability assumption, and occurrence of unexpected events affecting the outcome variable of one of the two groups but not the other, are the main threats to internal validity.

However, if a valid control group can be retrieved, the Non-equivalent Control Group Design is typically considered the strongest quasi-experimental design in terms of external validity.

Depending on the study context, data on outcomes and background characteristics for both groups of participants can be available either before and after the programme (pre-test and post-test) or post-test only. In the first case, the selection for treatment can be modelled on pre-test characteristics and the pre-test outcome can be compared to what happens after treatment is implemented (pre-test/post-test non-equivalent control group). Alternatively, when only post-test information is collected, we can rely on time invariant characteristics, collected after the intervention to model treatment group assignment (post-test-only non-equivalent control group). Several statistical techniques have been elaborated to analyse data from the non-equivalent control group designs and identify covariates to be included in statistical models that adjust for pre-test differences. Box 8 provides an overview of most common techniques. Box 9 indicates data requirements for implementation of different research designs.

Take-away message:

The evaluator's good understanding of the intervention design, i.e. the selection criteria for the programme, is a key element when selecting the most appropriate technique for evaluation design and analysis.

BOX 8. IMPLEMENTING QUASI-EXPERIMENTAL DESIGNS: DATA ANALYSIS METHODS

Difference-in-differences

The difference-in-differences method (DiD) compares changes in outcomes over time between units enrolled in the programme (the treatment group) and not (the comparison group). The change in outcome that takes place in the comparison group between the two periods is taken as a counterfactual of what would have happened to the treatment group in the absence of the intervention. Subtracting the change in the outcome observed in the comparison group from that observed in the treatment group gives the measure of impact (see Figure 6 in Appendix A).

Therefore, a key condition for DiD is to assume equal trends between the two groups in the absence of treatment. A good validity check for this is to compare changes in outcomes for the treatment and comparison groups repeatedly before the programme is implemented, when data are available. Alternatively, a placebo test performed on a group that we know was not affected by the programme, is another way to test the assumption. The DiD method allows us to get rid of the effects of all factors that do not change over time or that do not affect changes over time. As determinants of treatment participation can be expected to be time invariant in many contexts, this approach is very attractive for impact evaluations.

The great advantage of DiD is that treatment and comparison groups do not need to have the same conditions before the intervention roll-out. As a reminder, the model generates an average treatment effect on the treated (ATT), but it does not measure the effects of the intervention on the overall population. However, DiD attributes to the intervention any differences in trends between the two groups that occur at the time of the intervention. If any unaccounted factors affect this difference in trends, the estimation will be invalid or biased. DiD techniques can be applied to interrupted time series designs and wherever the parallel trend assumption is verified.

Synthetic Controls

The Synthetic Controls method is an extension of DiD where the parallel trends assumption is relaxed. The method uses information about the characteristics of the treated unit and the untreated units to construct a "synthetic" comparison unit by weighting each untreated unit in such a way that the synthetic comparison unit most closely resembles the one treated. In this way, covariates and outcomes of the control correspond to those of treatment prior to the intervention. In practice, a panel regression of outcomes on covariates (excluding treatment) is conducted, and a binary variable indicating the treatment status of individual observations is specified. By employing an optimization procedure, a weight is computed for each individual comparison group observations, such that the weighted synthetic control trends in covariates and outcomes match those of the treated units prior to treatment as closely as possible. The main advantages of synthetic controls are that the estimation of the treatment effect can be conducted even when the number of treated units is small; and bias is reduced when the "parallel trends" assumption underpinning DiD does not hold. However, this technique is not as efficient as DiD when the parallel trend assumption is valid.

Propensity score matching

Propensity score matching (PSM) methods generate control groups by matching treatment observations to one or more observations from the untreated sample, based on observable characteristics. The propensity score is the probability of ending up in the treatment group given observable characteristics. Treated units are matched with untreated units having a similar propensity score. Propensity scores are estimated with a probit or logit model, where the explanatory variables include all observed variables at the baseline, which may determine participation. For PSM estimates to be unbiased, average characteristics of the treatment and comparison groups need to be statistically identical prior to the intervention.

Several methods are available for conducting matching. Nearest neighbour matching matches the treatment unit to members of the control group with the nearest propensity score. More often the nearest five neighbours are chosen for matching. Alternatively, Caliper matching select comparison observations among those with propensity scores within a certain "distance" from propensity scores of treated units, and kernel matching includes all comparison observations in the region of common support with a weight that is inversely proportional to distance.[12] A single observation in the comparison group may be matched to several different observations in the treatment group. Members of the comparison group who do not match those treated are not considered in the estimation.

[12] The region of common support is identified by excluding observations in the untreated group with a propensity score lower than the lowest observed value in the treatment group, and observations in the treatment group with a propensity score higher than the highest observed value in the untreated group.

A key advantage of PSM is that it is always possible to use a binary treatment if there are sufficient data, and that it can be done *ex post*. If baseline data are not available, matching can be carried out on time invariant characteristics, such as sex and religion, or by recalling to pre-intervention characteristics, which can be retrieved at the endline, such as education of household head and ownership of major assets. PSM can generate both ATT and ATE. However, PSM approaches do not get rid of bias as a result of selection of unobservables. However, if selection into programmes is affected by unobservable characteristics, PSM will lead to biased estimates.

PSM can be combined with Difference-in-Differences to reduce the estimation bias due to unobserved characteristics (e.g. in non-equivalent control group designs) if baseline data are available. Matching combined with DiD takes care of any unobserved characteristics between the two groups that are constant over time.

Source: Authors' own elaboration.

Practical tips!

A list of studies (when available, focusing on school feeding evaluations) that provide examples of implementation of quasi-experimental data analysis techniques:

- Difference-in-Differences - Kazianga, H., de Walque, D. and Alderman, H. 2014. School feeding programmes, intra-household allocation and the nutrition of siblings: evidence from a randomized trial in rural Burkina Faso. *Journal of Development Economics, 106, 15-34.*
- Synthetic controls - Bouttell, J., Craig, P., Lewsey, J., Robinson, M. and Popham, F. 2018. Synthetic control methodology as a tool for evaluating population-level health interventions. *J Epidemiol Community Health, 72(8), 673-678.*
- Propensity score matching combined with Difference-in-Differences. Aurino, E., Tranchant, J. P., Diallo, A. S., Gelli, A. 2018. School feeding or general food distribution? quasi-experimental evidence on the educational impacts of emergency food assistance during conflict in Mali. (Also available at **https://www.unicef-irc.org/publications/pdf/WP-2018-04.pdf).**
- Regression discontinuity design - McEwan, P. J. 2013. The impact of Chile's school feeding programme on education outcomes. *Economics of Education Review, 32, 122-139.*

BOX 9. DATA REQUIREMENTS FOR IMPLEMENTATION OF DIFFERENT RESEARCH DESIGNS

Difference-in-Differences (DiD): Implementation of the method requires data on outcomes from both treatment and control groups at baseline and endline. Pre-intervention data on the outcomes are desirable to test the parallel trends assumption. If matching is to be used, then data for matching are also required.

Propensity score matching: PSM requires data from both the treated and untreated population, from which the comparison group is drawn, preferably before and after the intervention. Alternatively, when panel survey data are not available, matching procedures can rely on time invariant characteristics, e.g. sex, religion, or information collected at endline by recalling pre-intervention characteristics. Data should include characteristics of the unit of analysis that determine programme participation. The sample needs to be larger than the sample size suggested by simple power calculations, since observations outside the region of common support are discarded.

Regression Discontinuity Design (RDD): Data on the assignment variable are required, including information on how strictly the threshold rule has been applied. Data on the outcome indicators are needed for a large enough sample of those considered for the programme, including both those who were accepted and rejected. Details on other variables can be useful to verify balance across the threshold, preferably from a baseline survey. Administrative data can be used to a large extent with this design, reducing the need for data collection.

Source: Authors' own elaboration.

2.3 Designing impact evaluation of HGSF programmes

In order to pursue the evaluation rigorously, the evaluator should clearly have in mind the population groups that will be affected by the intervention before the evaluation starts. As discussed earlier, HGSF are integrated interventions combining school feeding and home-grown components (see Step 1 and Figure 3). They can be considered as multi-intervention programmes in which each intervention focus on a different population group (e.g. school-going children and local farmers). Consequently, impact estimates can be obtained for each group separately, reflecting the multiplicity of domains underpinning the programme. For instance, if evaluators are interested in estimating the impact of HGSF on nutrition, the eligible population will comprise all school-age children and their households (e.g. as there may be spillover on siblings or some other forms of intra-household redistribution of resources). On the other hand, when evaluation focuses on local agriculture development, local farmers represent the main target group. In line with the purpose of the note, we mostly focus on designing impact evaluations for the home-grown component of the programme. Nevertheless, some references to the school feeding component are provided in Box 12.

With regard to food producers, the population groups affected by the intervention would vary with the procurement model adopted for the food supply, as discussed in Section C. In decentralized procurement systems either single smallholder farmers or cooperatives (or farming associations) can

alternatively represent the unit of analysis based on the characteristics of the food supply chain. In centralized procurement schemes, the programme would also interest other actors involved in the food supply chain, e.g. traders and intermediaries. As already stated, for the sake of simplification, the remainder of the note focuses on suggesting practical evaluation strategies for the two procurement schemes described in the background section, i.e. decentralized versus centralized, while leaving other potential procurement models out of the main discussion.

2.4 Experimental evaluations

BOX 10. DEFINITION OF IMPERFECT COMPLIANCE, CROSSOVER AND SPILLOVER EFFECTS

Imperfect compliance and **crossover** are discrepancies between assigned treatment status and actual treatment.

- **Imperfect compliance** occurs when some units are assigned to the treatment group do not receive treatment, and conversely when some units assigned to the comparison group receive treatment. This can occur if units that are assigned to a programme choose not to participate, or some intended participants are excluded from the programme because of administrative or implementation errors. As noted earlier, imperfect compliance leads to differences between ATE and ATT.
- On the other hand, **crossover** occurs when units in the comparison group are mistakenly offered enrolment in the programme or manage to enrol. Crossover can happen if the eligibility criteria for assignment to treatment are not fully enforced, monitoring is poor or there is poor informational flow between the designers of the intervention and implementers. For more details on evaluating interventions in the presence of crossover see Linnemayr and Alderman (2011). Crossover is problematic for the evaluation as it affects the design, and may lead to biased estimates of treatment effects. Implementers and researchers should devote particular care to ensure crossover is minimized.

Spillover effects arise when an intervention affects a nonparticipant, and they can be either positive or negative. Nonparticipants can be interested in the intervention through several channels: externalities, as in the case of vaccination programmes where vaccinating children in a village decreases the probability that non-vaccinated inhabitants will succumb to the same diseases; social interactions with participants (e.g. through exchange of information provided by the programme, e.g. on health practices), and context or general equilibrium effects, when an intervention affects the behavioural or social norms, or results in some economic effects on other variables, within the given context. ATE may change in the presence of spillover on subjects that are both eligible and not eligible to the programme, potentially leading to wrong policy recommendations and to the neglect of mechanisms through which the intervention works.

Angelucci and Di Maro (2015) discuss these issues in depth, and elaborate suggestions as to how to design impact evaluations taking spillover effects into account.

Source: Authors' own elaboration.

Decentralized food procurement

As stated before, when procurement is decentralized, smallholder farmers living in school catchment areas are the main beneficiaries of the farming promotion component of the programme. In ideal conditions, RCT represents the preferable evaluation strategy. These conditions are: the evaluation design takes place jointly with programme design (or at least before the programme starts); there are potential units of treatment that have never been treated and whose treatment will take place later. School catchment areas or small administrative units containing them are the optimal randomization units to be randomly allocated either to the treatment or control group. School catchment areas are usually small and numerous enough to allow for randomized assignment of treatment. Alternatively, if a programme establishes a gradual phasing in, the control group can be constructed by randomizing the time in which participants are enrolled in the programme (step-wedge trial, see Box 7). In most cases, however, the evaluation design takes place when assignment rules have already been decided and communicated so that it is not possible to change them and randomly allocate the treatment.

A clustered design at school catchment area (or village) level would be preferred to randomization at individual farmer level, to minimize the risk of spillover and imperfect compliance with treatment assignment (see definition in Box 10) (Gertler *et al.*, 2016). The implementation of supply contracts between treated farmers and schools can produce spillover effects on the control group through multiple channels. For instance, an increased demand for agricultural inputs (seeds, fertilizers, irrigation, etc.) may translate into a rise in prices for the same inputs that would negatively affect other local farmers by augmenting production costs. On the other hand, an increase in the demand for food may raise the price of farm products, benefiting all local producers. The size of these effects may largely change based on the magnitude of the school food procurement demand as compared to overall local food production, as well as the structure of the productive sector, i.e. number of smallholder farmers, composition of farmers' production across different products and so on.

Imperfect compliance may occur when any farmer assigned to the control group participates in the programme. For instance, if the harvest is bad, farmers enrolled in the programme (treatment group) may decide to subcontract the production to other farmers who had been selected as part of the control group, so as to fulfil contractual supply requirements. Thus, control farmers would indirectly benefit from the increase in demand following HGSF procurement. In this context, randomization at the level of the single farmer would produce biased estimations, as crossover alters original treatment and comparison groups hampering treatment estimation. In many cases, not all farmers selected for treatment would decide to participate in the programme (imperfect programme uptake). Thus, we would expect the average treatment effect to be lower bounded.

Centralized food procurement

In the centralized food procurement model, the entire district is enrolled in the HGSF programme and all households with school-age children are eligible for school meals. Procurement also occurs centrally at the district level, so all farmers in the district are eligible to become suppliers for school feeding. In this setting, implementation of an experimental design is less straightforward. An RCT would require

the conduction of randomization at the district level and comparison of outcomes for farmers located in treated and control districts. A sufficient number of districts is required to implement this design in order to produce estimates that can capture statistical differences in the programme effects between treatment and control groups. Moreover, in order to compare farm production across districts we need to assume similarity among districts in terms of farm production systems. This assumption is not verified in many practical cases. Districts are often distant from each other and may vary in type of crops, agroecological zones, land distribution and so on. Therefore, implementation of an RCT is recommended only where districts are small and homogeneous or there are a sufficient number of districts to conduct the randomization.

2.5 Quasi-experimental methods

Decentralized food procurement

In some cases, RCTs cannot be implemented since the evaluator cannot control the intervention design from the very beginning,[13] or ethical, financial, political and operational factors may hamper random assignment. For instance, it may be the case that all schools in the area covered by the programme have been enrolled and consequently all smallholders have become eligible as food suppliers. When randomization is not feasible or desirable, quasi-experimental methods provide useful tools to create a comparable control group.

In most cases, farmers who decide to sell their products to schools differ from the others by a set of both observable characteristics (e.g. socio-demographics, asset level, access to credit, etc.) and unobservable (e.g. motivation, propensity to risk, etc.). The most common way to take these differences into account in the design phase is to use Non-equivalent Control Group Design and the related analytical methods presented in Section 2.2. A properly constructed comparison group should be as similar as possible to the counterfactual not at the individual level, but on average. Ideally, the two groups should be exposed to the same environmental factors, political and natural conditions over the time the intervention is administered.

As a general rule, a comparison group can be retrieved by selecting comparable farmer households from available datasets targeting agricultural populations similar to programme participants in comparable areas, and measuring the outcomes of interest for the evaluation among those populations. In the context of decentralized HGSF programmes, control group units can be selected from among eligible but non-beneficiary smallholders in the same school catchment areas, in case of incomplete coverage of the programme.[14] However, given the proximity to treated farmers, the estimates may be affected by the same spillover and crossover effects mentioned for randomized experiments. A similar

[13] Very often the evaluation design takes place when assignment rules have already been decided and communicated so that it is not possible to change them and randomly allocate the treatment.

[14] In this case the control group would comprise smallholder farmers that fulfil the criteria for being enrolled in the programme but they do not engage in it. In this case it would be relevant to have information on why they do not sell food products to schools.

bias may rise by including non-eligible households from the same areas in the comparison group. In this case, a further bias may arise from the fact that non-eligible farmers are likely to present differences in unobservable characteristics with respect to the eligible. Most likely, eligibility criteria established for the home-grown component can provide some indication of the nature of this bias, e.g. land size, product quality, productivity. The contamination risk related to spillover and crossover effects can be reduced by selecting eligible households in nearby or similar communities. However, this exposes the evaluator to the risk of having a control group that is not similar to the treated group, if selected communities are exposed to different environmental conditions. In this case, the eligibility criteria for programme enrolment of schools and relative catchment areas can drive the selection of comparison communities.

Since unobservable characteristics are not measurable, the best approach is to construct two groups and test whether they are statistically identical across observable characteristics. Hence, any differences observed between treatment and comparison farmers can be attributed to the programme itself, and not any other intervening factor or pre-existing characteristics that might be driving these differences (Jetha *et al.*, 2017).

When the evaluator can intervene at the *design stage* of the impact evaluation, non-probability sampling methods such as convenience, snowball, purposive, and quota sampling can provide alternative options to reconstruct a counterfactual group. See Section 3.1 for a full discussion of these sampling techniques. However, non-probability sampling methods do not generally yield representative samples. On impact evaluation, White and Raitzer (2017) recommend the use of non-probability sampling only when financial or logistical constraints make probability sampling unfeasible.

At the *analysis stage*, propensity score matching can be applied to match farmers who become school suppliers to one or more observations from the untreated sample that "look" very similar according to these observable characteristics, but did not receive the intervention (see Box 8). The observables used for matching can be either farmer characteristics or the estimated probability to participate in the programme given these characteristics. Propensity score matching can be better conducted when baseline data are available, since this allows for matching farmers with pre-intervention characteristics. Otherwise, we need to rely on time-invariant characteristics or re-called information collected at endline (see Box 9). However, matching techniques present an intrinsic limitation, as they assume that, once the impact of the observable characteristics has been considered, the outcome is independent of the observable characteristics. This means that every variable that impacts on both the outcome and the selection for HGSF programmes is observable (strong ignorability assumption). In reality, this assumption will rarely be fulfilled, and it may, therefore, be necessary to implement different techniques, such as DiD or RDD.

In our context of interest, DiD estimation allows us to get rid of the estimation bias caused by unobserved differences between farmers enrolled in the programme and not, by comparing the production outcomes (or other selected outcomes) over time between farmers enrolled in the programme and the comparison group. This implies the assumption that differences between HGSF farmers and controls are stable over time and that both groups are affected identically by common shocks, e.g. weather or price shocks, during the intervention period. Combining DiD estimations with PSM permits us to get rid of selection for time-invariant unobservables, by discarding the estimation

bias due to unobserved characteristics that are constant across time between the two groups and that would impact on both selection and outcomes. See Aurino *et al.* (2018) and Tranchant *et al.* (2018) for more details on these evaluation designs in the specific cases of quasi-experimental evaluations of school meal programmes. Keeping these general considerations in mind, the selection of the most appropriate estimation technique would be based on the assumptions about farmer characteristics determining enrolment into the programme and available data.

RDD may be applied when many farmers are present in school catchment areas and assignment to school feeding procurement is determined by some threshold criteria, e.g. some level of agricultural production, minimum quality standards, area of land. This would rarely occur in decentralized procurement models.

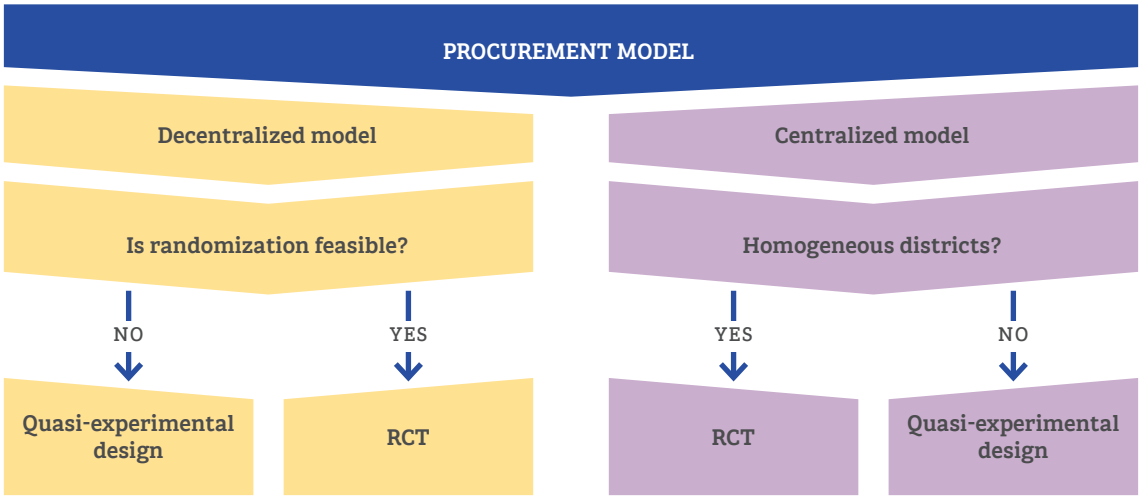
Centralized food procurement

Since the structure of the procurement system provides for all schools in the district to be treated, all farmers in their catchment areas are exposed to the programme, so they can potentially become school feeding suppliers. Randomized experiments are rarely possible at a single district level, since districts are often very different from each other in terms of farm production systems, e.g. type of crops, harvest conditions, land distribution, agroecology, etc. These elements expose the evaluator to the risk of having farmers in the control group that are not comparable to treatment groups since economic and geographical conditions in the district they live in are very heterogeneous.

Thus, quasi-experimental methods that identify the control group among farmers within the same district are preferable. Similar to what has been described for decentralized procurements, a first option would be to select controls from among eligible but non-beneficiary households in the same district, where coverage of the programme is incomplete. However, this entails risk of the spillover and crossover effects described above, with resulting estimation bias. Alternatively, non-eligible households from the same district can be included in the comparison group. However, farmers who sell their products to school feeding procurement may present different observable or unobservable characteristics with respect to the others. The choice of the preferred quasi-experimental design among those mentioned before would depend on the actual eligibility criteria and implementation rules defined by the specific HGSF programme.

In general, if the research team can control the evaluation design from the early design and implementation stages, matching procedures (propensity score matching and synthetic controls) can be more efficiently implemented. Baseline data collection can be carried out to improve matching quality by collecting a wide-variety of farmers' characteristics, reducing the estimation bias related to selection of unobservable characteristics. Availability of multiple rounds of data on key outcomes before intervention rollout would be ideal to test the equal trends assumption for DiD estimations (e.g. multiple waves of household surveys collected in the same country or region). However, this rarely occurs in practice, as it would require either retrieving available information from an existing survey or starting data collection long before the launch of the programme (at a high cost in terms of resources and time). As with the case of decentralized settings, if farmers' eligibility is determined by some threshold criteria, RDDs can be applied to estimate the impact of the intervention on farmers close to the threshold.

FIGURE 4. **SUMMARY OF DESIGN OPTIONS IN FACE OF DIFFERENT PROCUREMENT MODELS**



Source: Authors' own elaboration.

BOX 11. RESEARCH DESIGN WHEN FARMERS ARE ORGANIZED IN COOPERATIVES OR SMALLHOLDER ASSOCIATIONS

When selecting the evaluation design, evaluators should consider whether farmers were organized in cooperatives before the intervention, and whether they can sell their products only through cooperatives. This element will inform the choice of the most relevant sampling unit. Where there are no cooperatives, or farmers have no freedom of choice as to whether they sell their products directly or through cooperatives, research designs can consider farmers to be a single unit that can decide to participate in the programme. Otherwise, in contexts where farmers can sell their products only through cooperatives, these should be considered as treatment units instead, as all farmers who are members of treated cooperatives will enrol in the programme.

Implications for the research design depend on the food procurement models considered. For instance, a regression discontinuity design (RCT) design in a decentralized procurement model with all farmers in catchment areas selling through one cooperative would require the control group to be recruited among farmers outside that catchment area. In this case, as all farmers in treated clusters are participating in the programme, the RCTs would generate an estimation of the average treatment effect on the treated (ATT) in the programme instead of the *Average treatment effect* (ATE). Where many cooperatives are present in the same sampling unit, randomization at cooperative level can be strategic to investigate spillover effects on farmers who are not enrolled in the programme. In the centralized model, the treatment group includes all farmers who belong to cooperatives and decide to sell products to schools. In this case, farmers belonging to cooperatives that do not sell products to school can be included in the control group, and characteristics of the cooperatives (e.g. membership criteria, production and quality standards) can provide insights on drivers of selection into treatment. These considerations recommend that a careful preliminary investigation is made of farmer organizations before so as to shed light on the characteristics of food producers that can alter the research design. The same considerations should be made if programmes establish that farmers organize into cooperatives as part of the intervention.

Source: Authors' own elaboration.

BOX 12. **DESIGNING IMPACT EVALUATIONS FOR SCHOOL FEEDING INTERVENTIONS**

With regards to the impacts on nutrition and education outcomes, cluster RCTs allow for comparison of children in schools enrolled in the programme and not. Several empirical contributions can provide methodological insights on conducting evaluations of school feeding programmes on children by using a clustered randomized control design. Estimating the impact of school feeding on nutrition and education outcomes is more complicated if the intervention establishes that all school children in project areas are eligible for school meals. This would require selecting control children enrolled in schools from other districts. To do so, evaluators need to assume that districts are comparable, i.e. pupils and their households are similar and exposed to the same conditions (e.g. education system, food security conditions in the community, etc.). See for example Kazianga, de Walque and Alderman (2014); Buttenheim, Alderman, and Friedman (2011), Gelli *et al*., (2016).

Source: Authors' own elaboration.

2.6 Mixed-method process evaluation

Mixed method impact evaluation is based on a combined design comprising quantitative and qualitative analysis. Qualitative data collected employing a range of techniques, i.e. open and semi-open interviews with project staff and community leaders, focus groups, mapping exercises can be used at the formative stage to inform evaluation and survey design. More specifically, qualitative methods can help generate hypotheses and research questions before quantitative data collection, if conducted before quantitative survey design.

In addition, qualitative data can detect sensitive issues related to intervention of the implementation, such as barriers to participation, implementation problems, and so on. Sequencing, quantitative and qualitative data can be collected at the same time and used to triangulate findings or to generate early results on project impact. At a later stage, qualitative methods can support the interpretation of findings and provide insights to assess the external validity of the findings. Although information gathered during qualitative work cannot be generalized, it can capture tendencies, patterns and outliers. The information is useful in broadening understanding of the processes and pathways of impacts and their effects on different actors, both intended, unintended and unexpected (Creswell *et al*., 2003).

Take-away message:

Mixed methods provide an important contribution to researchers conducting impact evaluations of HGSF programmes for several reasons. At the initial stage, they can help deal with the complexities of the evaluation design by identifying key intervention steps and the range of beneficiary groups affected by the programme, i.e. local farmers, school children and their families, traders, cooperatives and farmer associations within the supply chain.

Information gathered using qualitative approaches provides important insights for quantitative survey design. It also provides rich and robust information that complement quantitative results, functioning to sharpen and expand understanding of impacts. In this regard, mixed methods help identify mechanisms through which the programme could increase its positive impacts, such as improvements in people's welfare and living standards.

Mixed methods also facilitate identification and explanation of why unattended spill-over effects have occurred among population groups targeted and not targeted by the intervention, including both positive and negative, as well as the measurement of non-economic outcomes, e.g. social networks, community cohesion, agroforestry knowledge and so on (Pozarny and Barrington, 2016; Borish *et al.*, 2017).

Qualitative research methods

The qualitative study, as part of a wider impact evaluation, can help unpack specific findings from a quantitative evaluation of HGSF programmes to explain and deepen understanding and the implications of findings, including, for example how institutional arrangements, design and implementation or operational processes lead to particular effects on household productive and consumption decisions. The qualitative research process requires varying degrees of flexibility to respond to contextual variation in each research region and community and to variation among interviewees and focus groups. The research must also be adapted to incorporate the local context.

A clear qualitative research process addresses the selection of research sites, selection of informants, timing and use of tools. Information is also provided with respect to ethical considerations, general behaviour, recording data and initial analysis. The research process in the approach adopted by FAO, for example, first consists of the preparation of a qualitative research design, articulated in a research guide which outlines: the overall roadmap of the research activity; the approach and methodology, including participatory tools; the guiding questionnaire with open-ended guiding questions; fieldwork protocol, site selection methodology; team member selection and training.[15]

[15] This paragraph and the one on sampling of research sites and informants for qualitative research were prepared by Pamela Pozarny, Senior Rural Sociologist, FAO, and Zahrah Nesbitt-Ahme

Fieldwork road map

The fieldwork "roadmap" is the first step in qualitative research design. The first step comprises conceptualization and concerns developing a set of key hypotheses to be tested in the field addressing priority areas of concern to the research. In this study, hypotheses were focused on impacts of HGSF on: income-generation; food and nutrition security; design, operational processes and institutional arrangements. These hypotheses are shaped around theories of change in this impact evaluation, particularly examining and understanding the pathways through which HGSF exerts influence on farm production, food security and educational outcomes, both directly and indirectly.

The roadmap then outlines the phases and steps in the overall field research process, including the sequencing of data collection process in each research community. For a comparative qualitative analysis, in each site, the research team split into two subteams working together in pairs, one as the facilitator and the other as note taker, they visit each main community (e.g. treatment and focus).

One key aspect of the approach is the visit to the comparison community (or non-treatment population, where the programme has not been implemented), which will be arranged in the research process near one of the key research communities. Locating a community that has not received assistance will need some pre-planning and coordination with implementing organizations. The members of the comparison community should have a similar, "comparative," socio-economic and agroecological profile context to the beneficiaries of the HGSF programme in the treatment communities. Note that the team works as one group on this day because time constraints preclude using a comparison community for both types of treatment community (remote and close); the team will need to decide whether to select a comparison community that is relatively far from or relatively near the main road and be able to justify this choice.

TABLE 2. **EXAMPLE OF FIELDWORK PROCESS ROADMAP FOR HGSF QUALITATIVE RESEARCH**

DAY 1	**Brief introduction at district level** **(KII* with informants from relevant ministries/and HGSF programme managers/officers)**	
	Village Cluster 1 (subteam 1)	**Village Cluster 2** (subteam 2)
	▸ Introduction to village Chiefs/leaders ▸ FGD** with men/women opinion leaders/ resource persons ▸ Mixed FGD with co-ops chairpersons benefiting from WFP P4P- with participatory tool ▸ KII with HGSF/WFP officer/programme implementers at field level ▸ Confirm fieldwork FGD/KII for next four days Evening debriefing	▸ Introduction to village Chiefs/leaders ▸ FGD with men/women opinion leaders/ resource persons ▸ Mixed FGD with co-ops chairpersons benefiting from WFP P4P- with participatory tool ▸ KII with HGSF/WFP officer/programme implementers at field level ▸ Confirm fieldwork FGD/KII for next four days Evening debriefing
DAY 2	▸ FGD with women HGSF beneficiaries - with participatory tool ▸ FGD with men HGSF beneficiaries - with participatory tool ▸ KII with leaders/presidents of cooperatives ▸ KII with agro-dealers benefiting from WFP P4P Evening debriefing	▸ FGD with women HGSF beneficiaries - with participatory tool ▸ FGD with men HGSF beneficiaries - with participatory tool ▸ KII with leaders/presidents of cooperatives ▸ KII with agro-dealers benefiting from WFP P4P Evening debriefing
DAY 3	▸ FGD with women HGSF beneficiaries - with participatory tool ▸ FGD with men HGSF beneficiaries ▸ KII with civil servants (e.g. health, agricultural extension) ▸ KII with leaders/presidents of cooperatives Evening debrief	▸ FGD with women HGSF beneficiaries - with participatory tool ▸ FGD with men HGSF beneficiaries ▸ KII with civil servants (e.g. health, agricultural extension) ▸ KII with leaders/presidents of cooperatives Evening debrief
DAY 4	▸ Household in-depth case study ▸ KII with teacher/head teacher ▸ KII that comes up during course of study (e.g. marketer) ▸ Brief community validation/feedback if time District Level feedback Evening debriefing	▸ Household in-depth case study ▸ KII with teacher/head teacher ▸ KII that comes up during course of study (e.g. marketer) ▸ Brief community validation/feedback if time District Level feedback Evening debriefing
Day 5	**Comparison community** ▸ FGD with men/women opinion leaders, using participatory tool ▸ FGD with female non-beneficiaries, using participatory tool ▸ FGD with male non-beneficiaries, using participatory tool	
DAY 6	**Team consolidation and synthesis half day**	

Source: Adaptation from Protection to Production (PtoP)/Oxford Policy Management (OPM) studies.
Note: The precise order of FGDs and KIIs may vary slightly between communities.
* Key informant interviews
** Focus group discussion

Debriefings

As a key part of the process, teams will start the initial data synthesis and analysis in the field. This begins at the level of the FGD or interview, with a check on data collected, but much occurs during the daily debriefing session.

Daily evening debriefings are a key stage of analysis in the research when the teams collectively reflect on and discuss their findings and analyse their working hypotheses from the day's fieldwork. The aim of this method of daily debriefings is to "build the story in the field" as the fieldwork transpires – adding to, contesting and strengthening findings and results to determine research hypotheses conclusions. The sessions also reveal knowledge gaps requiring follow up and further inquiry the next day. All team members "actively listen" and probe the researcher presenting during debriefings to sharpen information, gain greater clarity on initial summary findings, etc. It is essential that all team members participate actively in debriefings.

The output of these debriefings is a living field note document, organized around the research themes and related research questions, compiled by the lead country researcher, which will capture the key findings and gaps (under each area/questions) emerging from the discussion.

The daily debriefing sessions feed directly into a synthesis session conducted on the final day of fieldwork at each site, attended by all researchers. On the last day of synthesis, the findings are combined from all the sites and used to define the main study conclusions and to brainstorm preliminary recommendations. The team leader is responsible for leading and moderating the discussion during debriefs and synthesis exercises to systematically analyse, consolidate and synthesize the findings from all previous days of fieldwork to define main study conclusions and brainstorm preliminary recommendations.

Community feedback

At the end of the fieldwork in each community, each subteam carries out a community feedback session to report back to FGD participants and key informants on its preliminary findings. This feedback is essential as part of an ethical approach to research, and to validate findings and preliminary conclusions, offering community members an opportunity to add any last points. This session is critical to enabling ownership and sharing of the findings with the community. It also provides the subteam with the chance to validate its findings and preliminary conclusions, and to offer community members time to add any last observations.

Fieldwork

Prior to entering the field for data collection, according to the fieldwork protocol, the researchers contact the village head/chief for introductions in each community to explain the purpose of the study and request permission to undertake the study in the community. At entry, a courtesy visit should be conducted. Each focus group should bring together four to six participants (up to eight, if necessary) to discuss the proposed research areas. With exception to the FGDs with opinion leaders, during FGDs

the teams will employ the use of a participatory tool. Individual interviews will be conducted with relevant key informants during KIIs, including community leaders, extension agents, cooperatives, head teachers, and HGSF programme staff who have particular information and/or perceptions about the programme and its impacts on various stakeholders. The purpose is to elicit insights, information or examples, views and opinions of HGSF and CASU impacts from a wide variety of sources. Finally, in-depth household case studies are conducted with beneficiaries at their households following the structure of the question guide.

General conduct during fieldwork and ethical considerations

It is very important to ensure that the research conducted is both ethical and accurate, and this section sets out some general norms of behaviour when working in a research area.

Qualitative research is both a technical skill and art, requiring sensitivity and a keen openness to listening and skill in soliciting information. When conducting data collection to attain the most robust information, it is important to gain the trust and confidence of the informants. First, researchers should seek informed consent from informants, asking informants to answer questions openly, and ensuring confidentiality. In addition, the following should be considered:

- Community members and research participants must not feel offended or demeaned by anything researchers do, say or ask, or by the behaviour of researchers in their community. It is their community and they must be respected accordingly.
- Expectations of community members and research participants must not be raised by anything that is done or said during the research.
- Potential respondents must also feel under no explicit or implicit pressure to participate, either from the research team or from those who are asked to help identify participants (such as village heads, community elders or leaders, etc.).
- The research will be more accurate if participants see no reason or pressure to adjust their responses in a particular way and if they feel comfortable during the interview. This does not mean the researcher does not probe, triangulate and ask for examples and evidence.

The research being conducted may appear very strange to members of the community. It involves asking a number of personal questions and selecting many respondents at random. Even if this type of research has been conducted in the community before, it is likely that people may have questions. It is important to explain very clearly what is being done and why, and to respond to any questions about the research, from anyone who asks, patiently, clearly and honestly.

STEP 3. DESIGNING THE SAMPLING STRATEGY

3.1 Drawing a sample for impact evaluations: general recommendations

Sampling is commonly defined as the process of drawing units from a population of interest to estimate the characteristics of that population. Sampling is often necessary, as typically it is not possible to directly observe and measure outcomes for the entire population of interest (Gertler *et al.*, 2016). A rigorous sampling strategy for impact evaluation has to be designed in a way to ensure the sample is representative of the population of interest, and allowing for identification of a valid control or comparison group (White and Raitzer, 2017). This is crucial to ensuring the external validity of the findings of the impact evaluations.

To draw a representative sample a three-step procedure is usually taken:

- determine the population of interest;
- identify a sampling frame;
- draw as many units from the sampling frame as required by power calculations (see sections following).

The sampling frame is the most comprehensive list of units in the population of interest that can be put together (Martínez-Mesa *et al.*, 2016). **An adequate sampling frame is fundamental to ensure the results obtained from analysing a sample can be generalized to the entire population**. Otherwise, if the sampling frame does not exactly coincide with the population of interest, coverage bias may arise, compromising the external validity of the results for the population of interest (Gertler *et al.*, 2016). In the context of an impact evaluation, **the eligibility rules of the intervention indicate the most adequate procedure for drawing a sample.**

Probabilistic sampling methods are the most rigorous approach to drawing sampling from populations, as they assign a well-defined probability for each unit to be drawn. The main probabilistic sampling methods are:

- **Random sampling.** Every unit in the population has exactly the same probability of being drawn.
- **Stratified random sampling.** The population is divided into groups, and random sampling is performed within each group. Every unit in each group (or stratum) has the same probability of being drawn. Stratification allows for drawing inferences about outcomes at both the population level and within each group. Stratified sampling can also be applied when the research design requires an oversample of the specific subgroups in the population. Eligibility criteria for programme enrolment create the conditions for conducting stratification and oversample the population subgroup eligible for the intervention, e.g. farmers with the productive capacity to supply the demand from the school for food. Oversampling ensures that targeted groups are included in the sample, even when representing minorities or remote populations, as home-grown beneficiaries. Because stratification leads to oversampling of some groups relative to their shares in the population, the members of these groups have a higher probability of selection. Hence, a lower weight needs to be assigned to them than the rest of the population when estimating the intervention effect on the entire population. See Lance and Hattori (2016) about post-stratification re-weighting issues.

- **Cluster sampling.** Units are grouped in clusters, and a random sample of clusters is drawn. As a second stage, either all units in selected clusters are included in the sample or a number of units within the cluster are randomly drawn. This means that each cluster has a well-defined probability of being selected, and units within a selected cluster also have a well-defined probability of being drawn.
- **Stratified cluster sampling.** As above, units are grouped into clusters and first stage randomization occurs at cluster level. Stratification by certain characteristics can be implemented either at the cluster level (first stage) or within each cluster (second stage).

Non-probabilistic sampling methods are sampling techniques that gather sample units through a process that does not involve random selection, thus not all individuals in the population have an equal chance of being selected. The most common non-probability sampling methods are:

- **Volunteer sampling.** The members of the sample self-select to participate in the study.
- **Convenient sampling.** The researcher includes those participants who are easy or convenient to approach.
- **Purposive sampling.** The members of the samples are approached based on pre-defined criteria.

Quota sampling. This sampling method is used when the population is heterogeneous, and no element of the population matches all the characteristics of the predefined criteria. Homogeneous subgroups based on a few characteristics of the target population, e.g. gender, age, ethnicity, are formed. Criterion quota is established depending on the nature of the evaluation, then participants are non-randomly selected from each subgroup on the basis of these quota.

- **Snowball sampling**. One element of the population is approached at a time and then is asked to refer the investigator to the other elements of the population.

See Lance and Hattori (2016) for a broader and more detailed overview of these methods (**https://www.measureevaluation.org/resources/publications/ms-16-112**). However, it is likely these sampling techniques do not generate representative samples of the population of interest as a whole. Thus, the risk of incurring coverage bias is frequent and needs to be properly analysed and addressed.

When selecting the preferred sampling strategy, some elements need to be carefully considered. The sampling strategy should align with the evaluation design. Therefore, an RCT requires a cluster sample design. In quasi-experimental methods, a cost-efficient sampling could target eligible comparison groups. If the programme is expected to have important spillover effects on non-beneficiaries, then analysis of spillover effects depends upon non-beneficiaries being included in the sample. These non-beneficiaries are different from the comparison group, which is represented by non-beneficiaries who will not experience spillover (White and Raitzer, 2017).

3.2 Sampling for HGSF programmes

3.2.1 Definition of the population of interest for HGSF

In the context of HGSF programmes, a first fundamental challenge is the identification of the population of interest (Martínez-Mesa *et al.*, 2016). By definition, eligibility criteria for school feeding and home-grown components are different. School feeding usually covers all school-age children in a selected district or area and relative households. The home-grown component benefits a smaller share of the population represented by local farmers and their households. The food procurement model adopted by the programme delimits the population of interest for the home-grown component. For instance, in a decentralized model, the intervention would target local smallholders or farmer organizations. The introduction of specific eligibility criteria can further restrict the definition of the population of interest. Moreover, farmers who are able to enrol in the programme, and become school suppliers, are often limited to those with the production capacity to produce surplus and sell to schools. These correspond to a subsection of the entire farmer population, so they need to be oversampled to ensure appropriate coverage in both treatment and comparison groups.

Conversely, in a centralized procurement system, where HGSF only buy through aggregators, the farmers' status, as members of a cooperative, would define the reference population. As the intervention serves multiple beneficiary groups, e.g. school children and local farmers, all have to be representatively included in the sample. Thus, specific guidelines for sampling design should be identified based on the food procurement model used for the HGSF (decentralized versus centralized) and the selected evaluation method (experimental versus quasi experimental).

3.2.2 Experimental designs

In the case of a cluster of RCT designs, such as those proposed in Section C, the identification of the primary sampling unit depends on the food procurement model. In the decentralized model, a multi-stage cluster sampling procedure should be applied, and school catchment areas would be ideally considered as primary sampling units (PSU). As a first step, a subgroup of representative catchment areas is randomly selected and then randomly assigned to either treatment or control group.[16] Households within catchment areas can be considered as secondary sampling units (SSU). As eligible farmer households represent a subgroup of the entire household population, a stratified design should be implemented to specifically target this subgroup and oversample them within both treatment and control arms. A listing exercise that collects information on farm production, characteristics of the farmers and other data that can help identify eligible households should be conducted at an early stage to make this stratification possible. However, multiple sampling stages can be introduced in line with the research question. The Government of Ghana School Feeding Programme case study presented in Box 13 provides an example of study design that permits the comparison of two modalities of school feeding procurement by further stratifying randomization at the district level.

[16] See Figure 5 in Appendix for a visual representation of this step-wise procedure.

BOX 13. COMPLEX EVALUATION OF THE GOVERNMENT OF GHANA SCHOOL FEEDING PROGRAMME AS AN EXAMPLE OF SAMPLING DESIGN FOR HGSF

The trial focused on assessing programme effects at both child-level (education, health) and district-level (agriculture), as well as comparing the effects of the HGSF programme to the standard Ghana School Feeding Programme (GSFP) on agricultural outcomes. In order to impact on these different sets of outcomes, programme components were delivered at different administrative levels: the school feeding service, which was hypothesized to affect mostly child education and health, was designed to be delivered at the school level, while the agriculture-related activities were delivered at the district level, also affecting communities that would not be offered school feeding at the local school.

A complex study design was therefore required, which was achieved through a multiple-level randomization (see Figure in this box). The multi-level design compares child-level outcomes (e.g. education, health) between children assigned to school feeding (both GSFP and HGSF modalities) and control communities, and agriculture impacts of the HGSF pilot relative to the regular GSFP at district-level (as the agricultural activities were organized at the district- level).

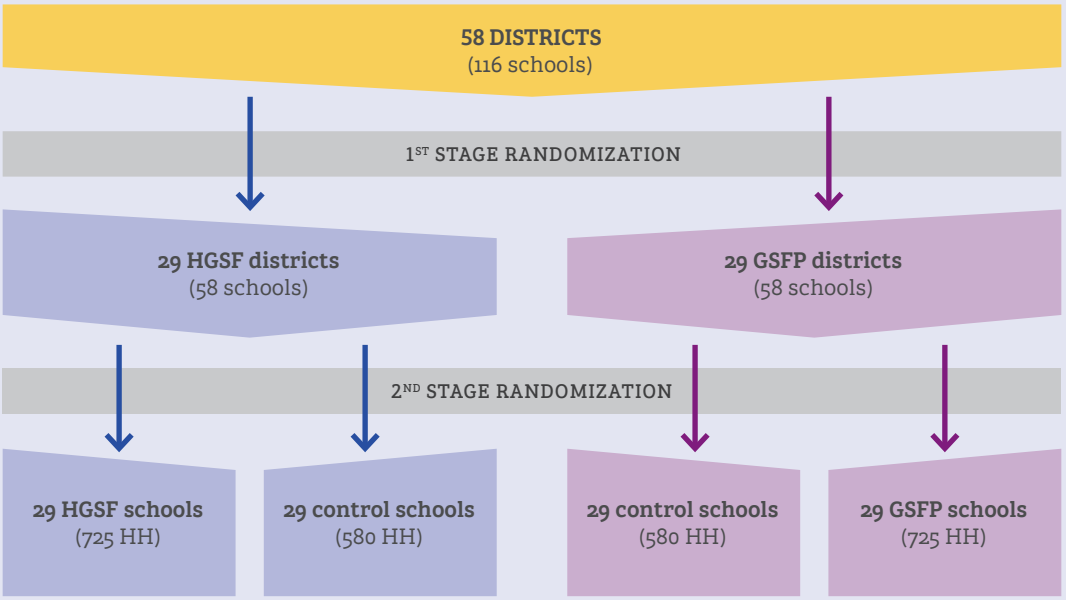
Identification of priority districts. To identify the population of interest a retargeting exercise was conducted to reach districts not previously covered by the programme in a sample frame that included all districts in the country. Poverty and food insecurity rankings were developed using the Ghana Living Standards Survey, the Core Welfare Indicators Questionnaire, the WFP Comprehensive Food Security and Vulnerability Assessment and other spatial data variables computed by the WFP. The data were then used to generate district-level composites for the share of national poverty and food insecurity. The sample frame was stratified by region, and 58 priority districts were identified with this exercise.

First-level of randomization, appropriate for estimating the differences between HGSF and GSFP with regards to agricultural outcomes: two comparable schools (and related catchment areas) were selected within each of the 58 selected districts and randomly assigned to HGSF or standard school feeding interventions. A list was obtained through the programme secretariat including schools not currently covered by the programme in each district. Data from the annual school census from 2011 to 2012 and Management Information System (EMIS) data were then used to match schools not receiving the GSFP and identify "best matched" pair.

Second-level randomization in order to estimate school feeding effects on child education and nutrition: Random allocation (lottery style) of schools to treatment and control groups by modelling pilot selection using a set of community and district-level variables and selecting the permutation of allocation that minimizes the regression best fit for the predicted selection.

Following a household census at baseline, power calculations and resource availability suggested the adoption of a sample of 25 households from the communities in the areas of the 58 schools receiving the intervention and of 20 households in the communities of the 58 control schools. Households were randomly selected from household listings in the catchment areas of the selected schools. The household listings were stratified into farmer/non-farmer households, based on agriculture classification data from the national census. Farmer households were sampled in both areas in the following way: 10 out of the 25 households in the 60 intervention communities were farmer households and five out of 20 households in 60 control communities were farmer households. Non-farmer households with children in the 5 to 15 years age group were randomly selected from the household listings.

This distribution of the sample between farmer and non-farmer households and between project control groups allows the construction of comparable samples. The multi-level design allows comparison of child-level outcomes (e.g. education, health) between children belonging to school feeding (including both GSFP and HGSF modalities) and control communities, and the agriculture impacts of the HGSF pilot relative to the regular GSFP at district-level.



Source: Aurino, Gelli, *et al.*, 2018.

Randomization of HGSF programmes in a centralized procurement setting is unlikely, however, theoretically possible (see discussion in Section C). PSUs should correspond to the randomization level. Thus, if the entire district is placed under the programme, districts should constitute the PSUs. Then households within selected districts (SSUs) can be sampled randomly or by introducing stratification layers, i.e. rural/urban, village size, and so on. In this setting, oversampling by strata of eligibility status is even more important since the share of those benefiting from home-grown is even smaller.

Where smallholders are organized into cooperatives or farmer associations, an additional sampling level should be considered to align the sample to the randomized design. As a general rule, when farmers are organized, the evaluation sample should accurately represent the population of eligible farmers for the results of the evaluation to be generalized and ensure external validity. Nonetheless, it is recommended that data be collected on farm production at the household level to estimate the actual impact of the programme on the ultimate beneficiaries, i.e. the local farmers.

3.2.3 Quasi-experimental designs

In quasi-experimental settings the sampling strategy should focus on targeting the eligible comparison groups. Programme eligibility criteria, e.g. production standard for supply farmers, help identify the population of reference from which the ultimate sample units are extracted. Administrative data can provide the list of all potential beneficiaries, reconstructing the sampling frame. As for randomized experiments, programme implementation modalities are determinant in designing the stratification strategy and identifying the PSUs, i.e. districts or school catchment areas based on the food procurement system adopted (see discussion in Section 3.2.2). Differently from the experimental context, researchers cannot randomize the treatment allocation of PSUs.

Depending on the procurement scheme researchers should identify the eligible populations for school feeding and home-grown components separately. Having this purpose in mind, they can decide whether all treated and control PSUs, where existing, or a random subgroup of them should be included in the sample to reach the targeted sample size. Households living in the selected PSUs constitute the SSUs and they should be randomly selected among the whole population eligible for the programme. Importantly for the home-grown component, both farmers who voluntary selected into the programme and not should be part of the sample.

Instead, for school feeding, since the entire population of school children can potentially be treated, controls should be randomly selected from children in non-treated PSUs. This is feasible only where control PSUs are comparable to the treated. Otherwise, if the whole eligible population is treated, sample units should be randomly identified among beneficiaries and their outcomes observed over time. As for experimental designs, the secondary sampling unit may change where farmers are organized into cooperatives. If many cooperatives are active in the study area, the researcher can decide to stratify the sample first at cooperative level and include all farmers in sampled cooperatives.

BOX 14. SAMPLING DESIGN FOR HGSF: PRACTICAL EXAMPLES

This Box illustrates a few examples to guide practitioners when identifying the most suitable sampling design for the evaluation of a hypothetical HGSF programme. We focus on estimating the effects of the home-grown component of the intervention on beneficiary farmers through the public food procurement mechanisms described in the introduction. In this hypothetical setting, the food procurement system is centralized at the district level, and the entire district is under the school feeding treatment. However, programme implementers have decided to focus only on a few subdistrict administrative units for the home-grown treatment. Within selected sites, only a few farmers provide food to the programme. Since the only eligibility criterion for farmers to be part of the home-grown programme is to reside in the district, the population of interest will be all resident farmers.

In this case study, evaluators know the home-grown treatment beneficiary status for all sites (clusters) in the subdistrict administrative units selected for the home-grown intervention, but not the treatment status of each farmer household. The intervention began before the evaluators could collect pre-intervention data. Hence, the only feasible evaluation design is a "post-test only non-equivalent control group" (Campbell and Stanley, 1963). A possible sampling strategy for this scenario would be a two-stage stratified cluster sampling with subdistrict administrative units as primary sampling units (PSU) and farm-households as secondary sampling units (SSU).

PSU should be first stratified by home-grown treatment status, then a random sample of PSU extracted. A listing exercise within selected PSU should be conducted to identify the treatment status of each single farmer household. Following, a second stratification exercise can be carried out at the household level to randomly extract a sample of farmers in both treatment and control PSU, where sample allocation to strata is disproportionate to allow for oversampling of rare populations, i.e. eligible farmers.

Now a similar scenario will be considered, whereby every cluster (PSU) has one farmer cooperative. Cooperative membership constitutes an eligibility criterion for the home-grown treatment. In this case, the second stratification would now depend on the level of analysis from which evaluators want to retrieve the effects of the programme. If they are interested in estimating the treatment effect on cooperative members, after PSU being first stratified by treatment status, a random sample of cooperative members in both treatment and control sites should be drawn. Alternatively, if they are interested in considering the implications of treatment for the overall population of farmers, a second stratification exercise should be carried out at the household level, by cooperative membership.

Source: Authors' own elaboration.

Sample size and power analysis

To understand the rationale behind the power analysis, first some definitions are required. The power of the study is defined as the probability that a study correctly detects the impact of an intervention that actually had an impact. The study comes to the right conclusion in two cases: when the intervention works and a significant impact is found, or it does not work and no significant impact is observed. When the intervention does not work, but the study concludes that it does, we have a Type I (inclusion) error. On the contrary, if the intervention does work but the study finds no significant impact, we end up with a Type II (exclusion) error. The power of a study is defined as 100 minus the probability of a Type II error.

A Type II error can be reduced by increasing the sample size. The objective of a power analysis is to determine the sample size the study needs to obtain for an acceptable level of Type II error. Practitioners usually consider an acceptable level as 20 percent, thus a power of 80 percent. In order to derive the sample size, we first need to establish the minimum detectable effect (MDE) – that is, how large (or small) an effect a study can detect. An MDE can be based on previous experiences of similar interventions or in consultation with policymakers. This may correspond to the policy objectives of the intervention. Statistically speaking, the MDE depends upon: the t–statistic values targeted for the significance level, usually corresponding to significance level threshold of 5 percent;, the chosen level of power (80 percent as default); standard error of the outcome variable;[17] the proportion of the sample in the treatment group and the sample size.

See calculations in Appendix B for both single and cluster designs. For further details on power calculations for standard and complex sampling designs see White and Raitzer (2017) and Gertler *et al.* (2016).

Take-away message:

Intuitively, the larger the MDE we want to detect, the smaller the required sample size. However, setting a MDE too large will result in an underpowered study, if the intervention does not have such a large impact as expected. The MDE is minimized with a "balanced sample", i.e. when we have the same number of observations in treatment and comparison groups. Power calculation software exists to calculate the required sample size on known parameters.

Importantly, when outcomes are highly correlated at the cluster level, e.g. income and test scores (large intra-class correlation coefficient), more power is achieved by sampling relatively more clusters rather than more people within cluster. This type of intervention requires samples with a large number of villages. A sample with many individuals from a relatively small number of villages will have limited power.

In addition, evaluators should consider issues related to subgroup analysis. As intervention effects need to be estimated in relation to small groups of farmers, we would like to oversample farmers in such a way to build groups that allow statistical tests to be performed. It is worth noting that variables such as income are measured with error and therefore come with large variance. This again requires larger samples.

[17] The value of outcome variables is unknown at the time of power calculations, as they need to be done before data collection. An estimate of outcome variables can be retrieved from other data source, preferably from the same country or context, i.e. administrative data, information collected during listing exercises.

Take-away message:

As a general criterion, the larger the sample, the more likely it is to be representative of the population from which the sample is taken. However, sample size has important cost/time trade-offs. Power calculations determine the minimum sample size sufficient for detecting statistically significant intervention effects in order to maximize the balance between representativeness of the reference population and the financial and opportunity costs of data collection.

A few operational tips follow: power calculations are a critical part of the study design and protocol power calculations have to be conducted separately for each outcome variable, and the largest required sample size for each outcome's power calculations will be the final sample size; power calculations should be independently checked by statistically skilled practitioners, and reaching a sufficient sample size is an essential condition to avoid investments in inconclusive studies (White and Raitzer, 2017).

Sampling of research sites and informants for qualitative research

The sampling design of the study sites for a qualitative research study should follow a consistent methodology across all study sites to strengthen the potential for comparative analysis and validity and reduce bias. A three-stage sampling process can be employed, although this can be modified based on the evaluation objective. This is explained in Box 14, which provides an adaptation of the sampling for a qualitative study undertaken for Zambia's home-grown school feeding programme.

Sampling regions

The will entail collaboration with the relevant programme-implementing agencies to sample two regions in each case study country for the fieldwork. The selection of these regions will reflect important differences in agroecological context, livelihoods and vulnerability. Or it may be purposive, should the programme agencies wish to explore in-depth features in particular contexts (e.g. highly productive zones, national border areas, high vulnerability locations).

Sampling districts

In each region, the qualitative fieldwork can be conducted in one district (or equivalent administrative area). For an evaluation of HGSF, the selected districts should be from the programme areas. In addition, it is envisaged that, for a mixed methods approach, at least one of the two districts chosen should be covered by the quantitative survey; this will maximize the opportunity for cross-fertilization of study results and the analytical potential of the mixed-method approach. The unit of analysis of the sampling and research activity will depend on the particular country's administrative organizational structure and the HGSF programme's implementation arrangements.

BOX 15. SAMPLING STRATEGY FOR IN-DEPTH QUALITATIVE PROCESS EVALUATION OF ZAMBIA'S HOME-GROWN SCHOOL FEEDING AND CONSERVATION AGRICULTURE SCALE UP PROGRAMMES

Following a quantitative impact evaluation of the home grown school feeding (HGSF) and the Conservation Agriculture Scale Up (CASU) programmes in Zambia, and in-depth qualitative study was undertaken to contextualise the findings. The qualitative study was based on a comparative analytical approach - focusing on farmers and cooperatives living in farming blocks covered by the HGSF and CASU programme compared to the farmers and cooperatives in HGSF alone. To capture the breadth of differences, each sample was examined in two districts: Luwingu (HGSF) and Katete (HGSF + CASU).

In order to probe in detail the findings of the quantitative impact evaluation, the qualitative study sampling paralleled the quantitative study, regarding two specific components:

- HGSF: households that benefit only from HGSF but not from CASU (i.e. households supplying to the World Food Programme's P4P programme where children receive school meals provided by the HGSF programme); and
- HGSF and CASU: farm households that benefit from CASU in districts where school meals and local procurement of pulses are available through the HGSF programme.

The sampling strategy involved a three-staged hierarchical approach of selecting districts, followed by sampling blocks and then selecting camps within each block. Additionally, the sampling strategy involved stratifying and sampling focus group participants within selected camps. The following methodology was used to select sites for fieldwork.

Site selection

Districts were the first level of selection for the study. The study districts mirrored that of the quantitative study. In the quantitative study, for the HGSF-only arm, the districts were selected and the survey concentrated on members of cooperatives who had benefited from the P4P and who lived in districts where school meals are provided under the HGSF programme. The identified district for the HGSF+CASU arm was made up of households that received support to conservation farming and live in farming blocks covered by the HGSF both in terms of public food procurement and school meals.

The second and third level of sampling was at the block and camp level. To capture the breadth of differences, each sample was examined in two blocks in Katete and two blocks in Luwingu, drawn from the list of blocks from the quantitative survey. A number of criteria were established for the selection of the block/village study locations, which included inter alia: overlap with the quantitative study; sufficient numbers of beneficiaries to conduct FGDs; and logistical feasibility.

Within each of the two selected blocks, the team selected those camps having a sufficient number of available beneficiaries to conduct research - a maximum of 16 male and female beneficiaries per camp. A camp with a low number of beneficiaries (particularly male beneficiaries) dictated the need to conduct research in more than one camp within the block. Drawing on support and partnership with CASU and HGSF staff in the combined site and HGSF alone site, this led to the selection of two to three camps where there were a sufficient number of beneficiaries for each block. The camps within these blocks then formed one study site.

Fieldwork sampling strategy

Districts	Type of site	Block	Camp
Katete	HGSF + CASU	Eastern Southern	Kamphambe Chilembwe

Source: Nesbitt-Ahmed and Pozarny, 2021.

STEP 4. MEASURING THE IMPACT ON MULTIPLE OUTCOMES

An HGSF programme normally combines the objectives of an SFP, e.g. education, nutrition or health outcomes – with additional goals related to the home-grown domain of the intervention, such as access and participation of smallholder farmers in a stable market. In the HGSF context, community participation and cohesion are additional elements that can contribute to the success or failure of a programme, and thus need to be evaluated. The design of indicators to be measured during the impact evaluation should build on the consideration of all potential interactions that occur among these stakeholders along the steps of the food supply chain, as discussed in Step 1 section. Furthermore, in mixed methods, the qualitative design fieldwork guide should be designed in parallel to the quantitative surveys, examining processes of interaction and capturing causal effects (intended and unintended) among different dimensions of the programme and stakeholders.

4.1 Measuring the impact on school children's education, nutrition and health

The impact on school children's education, nutrition and health has been largely measured by adopting a series of indicators for schooling outcomes (school enrolment, attendance, drop-out) and learning outcomes (indicators of cognition development and learning achievement). Gelli (2010) and Jomaa *et al.* (2011) present an extensive review of this evidence. Most of the attention on nutrition and health outcomes has been driven by the analysis of intervention effects on children anthropometric outcome (Alderman and Bundy, 2012; Jomaa *et al.*, 2011).

Practical tips!

For an overview of indicators for measuring the effectiveness of school feeding on nutritional and health status see FAO and WFP (2018). The document presents specific indicators for measuring schoolchildren's access to fresh and diverse food and dietary diversity.

Evaluating the impact of a programme on education outcomes requires the use of an established system to register and track schoolchildren's attendance and performance in class, at either school or household levels, or using both data sources. Nutritional and anthropometric outcomes need to be collected at the baseline and thereafter at regular intervals using measurements either at school or at home. Individual dietary diversity scores can be obtained through child-focused interviews (see Box 16).

BOX 16. MEASURING NUTRITIONAL OUTCOMES USING ANTHROPOMETRIC INDICATORS

A range of measures can be applied to assess child nutritional status. We review the most common below, with a particular focus on those used to measure nutritional status among children in the framework of Sustainable Development Goal 2.

Height-for-age z-scores (HAZ) assess the growth of children of a specific age and gender against a reference population. The **prevalence of stunting** captures how many children have HAZ below -2 standard deviations from the median of the reference population. This indicator, with reference to children younger than 5 years, is **Indicator 2.2.1 for the SDG2.**

Weight-for-age z-scores (BAZ) assess contemporaneous nutritional status, as compared to children of the same age and gender from the reference population. The **prevalence of wasting** assesses how many children have BAZ < -2 standard deviations from the median of the reference population, while the **prevalence of overweight** provides an assessment of the proportion of children having BAZ>1SD from the median of the reference population. Both indicators (with reference to children younger than 5 years) are included as key metrics to measure malnutrition in **Indicator 2.2.2 of SDG2**.

Source: Authors' own elaboration.

4.2 Measuring the impact of HGSF programmes on agricultural development

FAO and WFP (2018) provide a list of suggested outcomes, outputs and indicators specifically to evaluate the impact of HGSF programmes on farm and agricultural-related realms. General recommendations for evaluating the intervention impact on these outcomes are:

- all person-based indicators should be disaggregated by sex;
- trend analysis should be conducted in the short, medium and long term;
- representative samples of smallholder farmers should include both those who received support from the programme and those who did not;
- adoption of a survey system, based on telecommunications technology, e.g. using cell phones;
- avoid the inclusion of respondents in multiple surveys to boost households' willingness to participate and the quality of the answers obtained.

Increased farm production and market participation of smallholder farmers are considered core outputs of HGSF interventions in terms of food security and local agricultural development. FAO and WFP (2018) propose the following set of outcomes to evaluate HGSF impact on these dimensions. Three main outcomes are identified: degree of participation of smallholder farmers in the HGSF programme; effects of smallholder farmers' participation on their production and productivity; effects of increased production, productivity and market participation on smallholder farmers' vulnerability. Specific indicators are selected to measure each of these three core areas (Table 3).

TABLE 3. **SELECTED INDICATORS**

Outcomes	Indicators
Degree of participation of smallholder farmers in the HGSF programme	Volume and value of food purchased by the HGSF programme from smallholder farmers, by commodity
	Number of smallholder farmers who sold food to the HGSF programme
	Volume and value of sales from smallholder farmers to targeted aggregators
	Number of smallholder farmers who sold food to targeted aggregators
Effects of smallholder farmers' participation on their production and productivity	Number of farmers who increased their agricultural output, by commodity. By how much these farmers increased their output, by commodity.
	Number of farmers who increased their agricultural productivity (yield/ha), by commodity. By how much these farmers increased their productivity, by commodity.
	Number of farmers who diversified their agricultural production. By how much these farmers diversified their agricultural production, by commodity.
	Number of farmers who increased the profits of their agricultural production, by commodity. By how much these farmers increased the profits of their agricultural production, by commodity.
	Number of farmers who reduced post-harvest losses through improved techniques or participation in post-harvest handling and storage services. By how much these farmers reduced their post-harvest losses.
	Number of farmers who obtained access to credit for increasing their production and/or productivity. How much credit these farmers accessed for increasing their production/productivity.
Effects of increased production, productivity and market participation on smallholder farmers' vulnerability	Diversity of crops and animal products produced
	Dietary diversity score and food consumption score of smallholder farmers
	Coping strategy index of smallholder farmer households
	Share of expenditure spent on food by smallholder farmer households

Source: adapted from FAO and WFP, 2018.

In addition, information is required on household *income*, *consumption* and *assets* to assess to what extent participation in HGSF programmes translates into increased well-being for local farmers. Moreover, data on *prices* of agricultural inputs and products are fundamental to understanding spillover effects of the intervention on the entire community and potential general equilibrium effects.

To obtain this information, smallholder households should be interviewed at the baseline and thereafter at regular intervals. Frequency of data collection should be adjusted to the purchasing cycles of the programme. These can be by school term or by month, for example. Where different HGSF commodities have different seasons/agricultural cycles – for example, cereals and pulses have

one cropping season, and fresh vegetables, fruit, milk or egg another – data collection should be adjusted accordingly. The survey should include relevant questions as to whether they have received complementary support, e.g. access to credit and technological training, on price of physical inputs and general information should be asked about farm size, total yield for each crop, harvest use and post-harvest losses.

Price of agricultural products should be collected both at the household and community level considering seasonal price variations. The source of data to construct these indicators may vary depending on the procurement model. Information can be retrieved from household surveys when smallholder farmers provide food to schools directly. When food procurement takes place through an aggregator, such as cooperatives or farmer organizations, aggregators may be asked to share summary information on food quantities and prices, characteristics of farmers, and percentage of sales revenues that go to the farmers. Obtaining relevant data can be more difficult where schools or caterers obtain food through more centralized procedures. In these cases, as food is provided to schools by a commercial trader who collects products from many smallholders or cooperatives, the same price and output information should be collected at each level of the food supply chain, in order to identify the distribution of profits among different actors.

4.3 Data collection

Given the diversity of data described in the previous section, multiple survey instruments need to be designed for data collection. The following is a list of survey tools that can be adopted to collect information at different levels of observation. We recommend piloting of instruments to assess potential weaknesses, length of survey administration and other similar issues before starting with data collection for the full sample. Of course these instruments are all complemented with the qualitative research, and provide yet another instrument or method to collect information and triangulate findings.

Child survey: Individual questionnaire for school children including questions on schooling (enrolment, attendance, learning), food consumption, dietary diversity and health. Anthropometric indicators (height, weight) should be assessed by interviewers using standardized procedures. Data on school meals receipt can be collected at the child level as well (e.g. whether the child has consumed the meal in the previous week of schooling, how often the school meals are consumed, whether the meal is taken home to share with family members, whether less of other meals is eaten because of the school meal, etc.). For young children (5 to 6 years of age), the child's main carer may answer the questionnaire.

Household survey: The household questionnaire should target both school children and farmer households and collect information on household income and consumption, food security, housing and assets as well as socio-demographic indicators for each individual belonging to the household, e.g. age, gender, educational attainment, employment status and earnings, etc. Data on income and consumption should be collected at household level, since households may benefit from economies of scale, and report complementarities or substitutions in consumption. For example, housing expenses, cost of heating, water, electricity, food consumption cannot be fully captured from data at the individual level. However, relevant to this context, intra-household dynamics in the allocation of resources can

only be observed using data at the individual level within the household. The questionnaire should preferably include specific sections on participation of household members in programme activities, food and other agricultural prices, access to and cost of credit, engagement in community life and so on, to investigate contextual factors and behavioural mechanisms for programme impact.

Farmer survey: A questionnaire (or a module of the household questionnaire) specifically tailored to farmers should collect all information related to farm production: prices and purchases of agricultural inputs, food production for own consumption and sales, volume and value of food purchased by different buyers, livestock, post-harvest losses, access to credit and training, storage and shock exposure.

Community/Village survey: Information on commodity prices, infrastructure endowments, school facilities and exposure to covariate shocks such as droughts, pests, conflict, etc. Importantly for the context, a preliminary mapping of cooperative and farmer organizations operating in the area is fundamental to the selection of the research design and sampling.

School survey: collecting administrative data on enrolment and attendance, as well as SFP implementation.

Cooperative survey: Where cooperatives or farmers' associations are active in the study area, a specific survey should be designed to collect data on the cooperative's size and structure, eligibility criteria for membership, implications of cooperative membership for buying and selling products to the market including quantities and price standards, additional services provided to farmers.

In the context of HGSF programmes, practitioners are often interested in testing the effectiveness of an intervention on multiple outcomes, for multiple subgroups, at multiple points in time, or across multiple treatment groups. This means that a number of statistical hypothesis tests are simultaneously specified, which can lead to spurious findings of effects. Multiple hypothesis testing methods should be applied in these circumstances.

Indeed, when testing a single hypothesis, researchers typically specify an acceptable maximum probability of making a Type I error, . A Type I error is the probability of erroneously rejecting the null hypothesis when it is true, and it is usually set at 0.05. Consequently, when testing multiple hypotheses by conducting a separate test for each of the hypotheses, the overall probability of a false positive finding in the study becomes greater than 5 percent. This Type I error inflation is maximum for independent outcomes. Multiple outcomes are usually somewhat correlated, which correlates the test statistics and reduces the extent of Type I error inflation. However, any error inflation can still be problematic when drawing reliable conclusions about the existence of effects (Schochet, Feldman and Burghardt, 2008).

Multiple testing is a statistical procedure that deals with this problem by adjusting p-values for the effect of Type I error inflation. A relevant implication of employing multiple testing is a reduction in statistical power. Power of an individual hypothesis test is the probability of rejecting a false null hypothesis of at least a specified size. If p-values are adjusted upward, one is less likely to reject the null hypotheses that are true, which reduces the probability of Type I errors, or false positive findings. At the same time however, we are also less likely to reject the null hypotheses that are false. Therefore, multiple hypothesis testing reduces the power of each separate hypothesis compared with the situation when no multiplicity adjustments are made (Porter, 2016).

BOX 17. CHALLENGES IN TESTING MULTIPLE OUTCOMES: MULTIPLE HYPOTHESIS TESTING

Bearing this trade-off in mind, the multiple testing strategy should be based on a process that first groups and prioritizes selected outcomes since the design stage of the study, then orient data collection and analysis at later stages. Multiple comparison corrections should not be applied blindly to all outcomes, subgroups, and treatment alternatives to avoid large reductions in the statistical power of the tests. Selection of outcome domains to be tested should build on the theory of change proposed for the intervention. A domain can be defined by grouping outcomes with a common latent structure (such as test scores for different skills, behavioural outcomes, etc.) or grouping outcomes with high correlations. Outcomes will likely be grouped into a domain if they are expected to measure a common latent construct. Thus, conducting tests for domain outcomes as a group will measure intervention effects on this common construct. A domain can entail a same outcome measured over time, or outcomes referring to a specific population subgroup (Schochet *et al*., 2008).

Most multiple hypothesis testing conducted for impact evaluation belongs to two different classes. The first group of techniques usually set a family wise error rate (FWER), i.e. a Type I error rate across the entire set of outcomes (Bonferroni, Holm, Westfall and Young). In other words, these multiple testing methods adjust p-values in a way that ensures that the probability of at least one Type I error across the entire set of hypothesis tests is no more than a certain percentage, usually 5 percent. The second class of multiple testing procedures (MTP) takes an entirely different approach to the multiple testing problem. MTPs in this class control for the false discovery rate (Benjamini and Hochberg, 1995). The FDR is the expected proportion of all rejected hypotheses that are erroneously rejected.

Source: Authors' own elaboration.

Practical tips!

Several statistical approaches have been developed to run multiple hypothesis testing. For more details see: Benjamini and Hochberg (1995); List *et al.* (2019); Benjamini, Krieger and Yekutieli (2006).

STEP 5. CONSIDERING IMPLICATIONS FOR EXTERNAL VALIDITY

We refer to external validity of the intervention as the ability to generalize the results of the programme and to transfer them to other contexts that differ from the context of intervention. The common threats to external validity, linked to the characteristics of the research design selected to evaluate the impact of the intervention, have been discussed for both experimental (Section 2.2.1) and quasi-experimental (Section 2.2.2) designs. Some practical examples of common external validity threats in the context of HGSF are given in Sections 2.4 and 2.5.

However, HGSF interventions are implemented using a variety of procurement models and supporting factors in such a way that simple generalizations from one context to the next may not often be possible. The question "does HGSF work" is not very informative and unlikely to be answered by any study or number of studies.

The success of the intervention will depend on the prevailing characteristics of programme implementation, such as: programme design; target group characteristics; implementing agency characteristics (e.g. administrative, monitoring and financial capacity, and whether these vary in specific areas where the programme is implemented); geographic scale of the programme and heterogeneity in the areas affected by the programme (e.g. in terms of agroecology, ethnicity, production capacity, access to credit, and so on); market structures and prices; contextual institutional setups, including levels of trust between institutions and HGSF actors, as well as between farmers and schools, etc.[18]

Given these heterogeneities, HGSF may work in some contexts, but not in others. It may work for some farmers, but not for others, and so on. Can the results of an impact evaluation of HGSF intervention be extrapolated from one context to another? No, this is unlikely to be feasible, unless the contexts are extremely similar, which in most instances will not be the case.

- *What can evaluators and researchers possibly do to say whether findings of their evaluation apply to different settings or countries?* In our view, the focus of the evaluators and policymakers should not be on an acritical extrapolation of findings to new settings, rather on understanding the mechanisms that made the intervention work, or fail to work. For example, our theory of change analysis has identified some key general questions that are preconditions for the successful operation of HGSF across contexts: does HGSF generate additional food demand in the market? Do farmers respond to the additional demand by producing more food? Will farmers make investments in their farm to respond to larger and more stable demand for food?
- It seems that many of the expected benefits of HGSF depend on a positive answer to these questions. Note that these questions are rather general in formulation and refer to the fundamental mechanisms that allow for the operation of HGSF. The structure of the markets may vary across countries, as well as the characteristics of the farmers, but we would expect a positive answer to the questions above for the programme to be successful.
- Simple extrapolation is not possible because the success of HGSF is not simply dependent on a set of given characteristics of the population or of the intervention, such as for example, the type of local food being produced or the type of contract between farmers and the project. The success of the project also depends on several supporting factors including the presence of simultaneous programmes and institutional conditions allowing the operation of agricultural markets and contractual arrangements. A successful replication of the intervention would require the presence of the same or similar contextual factors. Hence, an analysis of the external validity of HGSF should consider the key factors that allow the operation of the fundamental mechanisms described above.

[18] In particular, programme scale and identity of programme implementer (e.g. international organizations versus government) matter for external validity: see discussion in Aurino et al. (2018), for a list of useful references on the generalization of evaluation results based on small-scale school feeding programmes implemented by international agencies.

Concluding remarks

HGSF initiatives are widespread modalities that are employed to combine the provision of school feeding with agricultural development objectives, using food produced and purchased within the country. The main objective of linking school feeding to agriculture development – particularly to local small-scale production – is to reduce rural poverty by developing markets and generating a regular and reliable source of income for smallholder farmers. Although HGSF initiatives have been implemented in many countries, empirical evidence to assess the effectiveness and economic sustainability of such programmes with regard to agricultural goals is limited.

This guide provides practical tools for conducting rigorous mixed method impact evaluation of HGSF initiatives and addresses the main methodological challenges of measuring the real impact of programmes through a stepwise approach. The focus is placed mostly on the agricultural development component of the programmes, as this area is where the largest knowledge gaps remain. To tailor the guide to common implementation modalities of HGSF, two specific frameworks are considered. First, a decentralized food procurement system in which the food-supply chain links local smallholders in school catchment areas directly to schools. In a second scenario, procurement is centralized and the food supply chain involves farmers in a larger area.

The guide analyses all stages of setting up a mixed methods impact evaluation of HGSF programmes. First, it discusses the setting up of a theory of change, and qualitative analysis – proposed hypotheses or areas of inquiry – by identifying the channels through which the purchase of school meals from local farmers can increase agricultural profits, and in turn farmers' incomes. Following, methodological issues are presented that concern the choice of rigorous research designs and sampling strategies, regarding both quantitative and qualitative research designs. Measurement of the intervention impacts on different dimensions is discussed, focussing specifically on agricultural production and farmers' income. Finally, there is discussion of the external validity issues concerning extrapolation of the results of a single intervention to other contexts.

References

Afridi, F. 2010. Child welfare programmes and child nutrition: Evidence from a mandated school meal programme in India. *Journal of Development Economics*, *92*(2), 152–165. https://doi.org/10.1016/J.JDEVECO.2009.02.002.

Afridi, F. 2011. The Impact of School Meals on School Participation: Evidence from Rural India. *Journal of Development Studies*, *47*(11), 1636–1656. https://doi.org/10.1080/00220388.2010.514330

Alderman, H. & Bundy, D. 2012. School feeding programmes and development: Are we framing the question correctly? *World Bank Research Observer*, *27*(2), 204–221. https://doi.org/10.1093/wbro/lkr005.

Angelucci, M. & Di Maro, V. 2015. *Programme evaluation and spillover effects* (Policy Research Working Paper No. 7243). Washington, DC, World Bank. Retrieved from http://econ.worldbank.org.

Arsenault, J. E., Mora-Plazas, M., Forero, Y., López-Arana, S., Marín, C., Baylin, A. & Villamor, E. 2009. Provision of a school snack is associated with Vitamin B-12 status, linear growth, and morbidity in children from Bogotá, Colombia. *The Journal of Nutrition*, *139*(9), 1744–1750. https://doi.org/10.3945/jn.109.108662.

Aurino, E., Gelli, A., Adamba, C., Osei-Akoto, I. & Alderman, H. 2018. *Food for thought? Experimental evidence on the learning impacts of a large-scale school feeding programme in Ghana.* IFPRI DIscussion Papers. Washington, DC, International Food Policy Research Institute.

Aurino, E., Tranchant, J.-P., Diallo, A.S. & Gelli, A. 2018. *School feeding or general food distribution? Quasi-experimental evidence on the educational impacts of emergency food assistance during conflict in Mali* (Innocenti Working Paper). Florence, Italy.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Benjamini, Y., Krieger, A. M. & Yekutieli, D. 2006. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491.

Borish, D., King, N. & Dewey, C. 2017. Enhanced community capital from primary school feeding and agroforestry programme in Kenya. *International Journal of Educational Development*, *52*, 10–18. https://doi.org/10.1016/j.ijedudev.2016.10.005.

Blamey, A., Mackenzie, M. 2007. Theories of change and realistic evaluation: peas in a pod or apples and oranges? *Evaluation*. 2007;13(4):439-455. doi:10.1177/1356389007082129.

Bundy, Donald, Burbano, C., Grosh, M., Gelli, A., Jukes, M. & Drake, L. 2009. 02-Rethinking school feeding: social safety nets, child development, and the education sector. *Human Development*. https://doi.org/10.1596/978-0-8213-7974-5.

Burde, D., Kelcey, J., Al-abbadi, K., Anastacio, A., Anderson, R., Balfour-poole, C., … Mckinney, R. 2015. *What works to promote children's educational access, quality of learning, and wellbeing in crisis-affected contexts.* Education Rigorous Literature Review. (Also available at https://www.edu-links.org/sites/default/files/media/file/Education-emergencies-rigorous-review-2015-10.pdf).

Buttenheim, A., Alderman, H. & Friedman, J. 2011. Impact evaluation of school feeding programmemes in Lao People's Democratic Republic. *Journal of Development Effectiveness*, *3*(4), 520–542. https://doi.org/10.1080/19439342.2011.634511

Campbell, D.T. & Stanley, J.C. 1963. *Experimental and quasi-experimental designs for research*. Boston, USA, Houghton Mifflin Company. (Retrieved from https://www.sfu.ca/~palys/Campbell&Stanley-1959-Exptl&QuasiExptlDesignsForResearch.pdf).

Chakraborty, T. & Jayaraman, R. 2016. *School feeding and learning achievement: evidence from india's midday meal programme.* Munich, Germany, Center of Economic Studies. (CESifo Working Paper Series No. 5994). (Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2821476).

Collier, P. & Dercon, S. 2009. Expert Meeting on how to feed the world in 2050 African agriculture in 50 years: smallholders in a rapidly changing world? *In Expert meeting on how to feed the world in 2050*. (Retrieved from http://ss.rrojasdatabank.info/ak983e00.pdf).

Creswell, J., Clark, V.L.P., Gutmann, M.L. & Hanson, W.E. 2003. Advanced mixed methods research designs. *In* Tashakkori, A. & Teddlie, C. (Eds.) (p. 768). USA, SAGE Publications. (Retrieved from https://books.google.it/books?hl=it&lr=&id=F8BFOM8DCKoC&oi=fnd&pg=PA209&dq=creswell+mixed+methods&ots=gVeUvDuxOf&sig=Ek-so7nIh0ZdvUfQUXVbvqclWXM#v=onepage&q=creswell mixed methods&f=false).

Davis, B., Di Giuseppe, S. & Zezza, A. 2017. Are African households (not) leaving agriculture? Patterns of households' income sources in rural Sub-Saharan Africa. *Food Policy*, *67*, 153–174. https://doi.org/10.1016/J.FOODPOL.2016.09.018

Devereux, S. 2015. *Home-grown school feeding and social protection*. (Imperial College, London, Partnership for Child Development, Working Paper No. 216).

Drake, L., Fernandes, M., Aurino, E., Kiamba, J., Giyose, B., Burbano, C., … Gelli, A. 2017. School feeding programmes in middle childhood and adolescence. *In* Bundy, D., De Silva, N., Horton, S., Jamison, D. & Patton G.C. (Eds.), *Disease Control Priorities 3*. Washington, DC, World Bank.

Drake, L., Woolnough, A., Burbano, C. & Bundy, D. (Eds.). 2016. *Global school feeding sourcebook lessons from 14 countries* (Vol. 91). London, Imperial College Press.

Duflo, E., Glennerster, R. & Kremer, M. 2007. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4, 3895-3962.

FAO. 2014. *School Feeding and possibilities for direct purchases from family farming*. (Retrieved from http://www.fao.org/3/a-i3413e.pdf).

FAO & WFP. 2018. *Homegrown School Feeding. Resource Framework*. Rome.

Gelli, A. 2010. *Food provision in schools in low and middle income countries : developing an evidenced based programme* (HGSF Working Paper Series No. 4).

Gelli, A., Aurino, E., Folson, G., Arhinful, D., Adamba, C., Osei-Akoto, I., … Alderman, H. 2019. The national school meals programme in Ghana improves height-for-age during mid-childhood, in girls and in children from poor households: A cluster randomised trial. *Journal of Nutrition*.

Gelli, A., Masset, E., Folson, G., Kusi, A., Arhinful, D. K., Asante, F., … Drake, L. 2016. Evaluation of alternative school feeding models on nutrition, education, agriculture and other social outcomes in Ghana: Rationale, randomised design and baseline data. *Trials*, *17*(1). https://doi.org/10.1186/s13063-015-1116-0

Gelli, Aulo, Hawkes, C., Donovan, J., Harris, J., Allen, S., De Brauw, A., … Ryckembusch, D. 2015. Value Chains and Nutrition CGIAR Research Programme on Agriculture for Nutrition and Health is a senior research fellow for the CGIAR Research Programme on Agriculture for Nutrition and Health at, (January).

Gelli, Aulo, Kretschmer, A., Molinas, L. & Regnault de la Mothe, M. 2012. *A comparison of supply chains for school food: Exploring operational trade-offs across implementation models.* (HGSF Working Paper Series No. 7).

Gelli, Aulo, Masset, E., Folson, G., Kusi, A., Arhinful, D.K., Asante, F., … Drake, L. 2016. Evaluation of alternative school feeding models on nutrition, education, agriculture and other social outcomes in Ghana: Rationale, randomised design and baseline data. *Trials*, *17*(1). https://doi.org/10.1186/s13063-015-1116-0.

Gertler, P., Martinez, S., Premand, P., Rawlings, L.B. & Vermeersch, C. M. J. 2016. *Impact Evaulation in Practice* (Second). World Bank Group. https://doi.org/10.1109/WI-IATW.2006.145.

Gockel, R., Hampton, K., Mary, P. & Gugerty, K. 2009. Storage , Transportation and Aggregation of Agricultural products for smallholder farmers in sub-Saharan Africa. *Organization*, 1–33.

Grantham-McGregor, S.M., Chang, S. & Walker, S. P. 1998. Evaluation of school feeding programmes: some Jamaican examples. *The American Journal of Clinical Nutrition*, *67*(4), 785S-789S.

Jomaa, L.H., McDonnell, E. & Probart, C. 2011. School feeding programmes in developing countries: Impacts on children's health and educational outcomes. *Nutrition Reviews*, *69*(2), 83–98. https://doi.org/10.1111/j.1753-4887.2010.00369.x.

Kazianga, H., de Walque, D. & Alderman, H. 2009. *Educational and health impacts of two school feeding schemes evidence from a randomized trial in rural Burkina Faso*. Washington, DC, World Bank Policy Research Working Paper No. 4976.

Kazianga, H., de Walque, D. & Alderman, H. 2014. School feeding programmes, intrahousehold allocation and the nutrition of siblings: Evidence from a randomized trial in rural Burkina Faso. *Journal of Development Economics*, *106*, 15–34. https://doi.org/10.1016/J.JDEVECO.2013.08.007.

Kristajnsson, E.A., Gelli, A., Welch, V., Greenhalgh, T., Liberato, S., Francis, D. and Espejo, F. 2016. Costs, and cost-outcome of school feeding programmes and feeding programmes for young children. Evidence and recommendations, *International Journal of Educational Development*, 48, p.79-83. (Also available at https://www.sciencedirect.com/science/article/pii/S0738059315300134).

Kretschmer, A., Spinler, S. & Van Wassenhove, L.N. 2014. A school feeding supply chain framework: Critical factors for sustainable programme design. *Production and Operations Management*, *23*(6), 990–1001. https://doi.org/10.1111/poms.12109.

Kristjansson, B., Petticrew, M., MacDonald, B., Krasevec, J., Janzen, L., Greenhalgh, T., … Welch, V. 2007. *School feeding for improving the physical and psychosocial health of disadvantaged students*. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD004676.pub2.

Lance, P. & Hattori, A. 2016. *Sampling and evaluation – A Guide to sampling for programme impact evaluation — MEASURE Evaluation*. Retrieved from https://www.measureevaluation.org/resources/publications/ms-16-112.

Lawson, T. M. 2012. *Impact of school feeding programmes on educational, nutritional, and agricultural development goals: a systematic review of literature* (Graduate Research Master's Degree Plan B Papers No. 142466). *Digital Times* (Vol. Unknown). (Retrieved from http://www.dt.co.kr/contents.html?article_no=2012071302010531749001).

Linnemayr, S. & Alderman, H. 2011). Almost random: Evaluating a large-scale randomized nutrition programme in the presence of crossover. *Journal of Development Economics*, *96*(1), 106–114. https://doi.org/10.1016/J.JDEVECO.2010.06.002.

List, J. A., Shaikh, A. M. & Xu, Y. 2019. Multiple hypothesis testing in experimental economics. *Experimental Economics*, 1–21. https://doi.org/10.1007/s10683-018-09597-5.

Ludwig, J., J Kling, J.R. & Mullainathan, S. 2011. Mechanism Experiments and Policy Evaluations. *Journal of Economic Perspectives*, 25 (3): 17-38. DOI: 10.1257/jep.25.3.17

Martínez-Mesa, J., González-Chica, D.A., Duquia, R.P., Bonamigo, R.R. & Bastos, J. L. 2016. Sampling: how to select participants in my research study? *Anais Brasileiros de Dermatologia*, *91*(3), 326–330. https://doi.org/10.1590/abd1806-4841.20165254.

Masset, E., Acharya, A., Barnett, C. & Tony Dogbe, T. 2013. An impact evaluation design for the Millennium Villages Project in Northern Ghana, *Journal of Development Effectiveness*, 5:2, 137-157, DOI: 10.1080/19439342.2013.790914.

Nesbitt-Ahmed, Z. and Pozarny, P. 2021. *Qualitative research on impacts of the Zambia Home Grown School Feeding and Conservation Agriculture Scale Up programmes*. Rome, FAO. https://doi.org/10.4060/cb4442en

Porter, K. E. 2016. *Statistical power in evaluations that investigate effects on multiple outcomes a guide for researchers*. (Retrieved from www.mdrc.org).

Poulton, C., Kydd, J. & Dorward, A. 2006. Overcoming Market Constraints on Pro-Poor Agricultural Growth in Sub-Saharan Africa. *Development Policy Review*, *24*(3), 243–277. https://doi.org/10.1111/j.1467-7679.2006.00324.x.

Pozarny, P. & Barrington, C. 2016. Qualitative methods in impact evaluations of cash transfer programmes in the transfer project in sub-Saharan Africa. In Davis, B., Handa, S., Hypher, N., Winder Rossi, N., Winters, P. & Yablonski, J. (Eds.), *From evidence to action: the story of cash transfers and impact evaluation in sub-Saharan Afric*a (pp. 71-93). Oxford, England; Rome, FAO.

Read, K.L., Kendall, P.C., Carper, M. M. & Rausch, J. R. 2013. Statistical methods for use in the analysis of randomized clinical trials utilizing a pretreatment, posttreatment, follow-up (PPF) paradigm. *The Oxford handbook of research strategies for clinical psychology*, 253-260.

Reis, H.T. & Judd, C.M. (Eds.). 2013. *Handbook of research methods in social and personality psychology*. New York, Cambridge University Press. https://doi.org/10.1017/CBO9780511996481

Rothman, K.J. & Greenland, S. 2005. Hill's Criteria for Causality. *Encyclopedia of biostatistics*, 4. (Also available at https://www.rtihs.org/sites/default/files/26902%20Rothman%201998%20The%20encyclopedia%20of%20biostatistics.pdf).

Schochet, P.Z., Feldman, A. & Burghardt, J. 2008. *Guidelines for multiple testing in experimental evaluations of educational interventions.* Princeton, NJ, USA, Mathematica Policy Research, Inc.

Singh, A., Park, A. & Dercon, S. 2014. School meals as a safety net: An evaluation of the midday meal scheme in India. *Economic Development and Cultural Change*, *62*(2), 275–306. https://doi.org/10.1086/674097.

Snilstveit, B., Stevenson, J., Menon, R., Phillips D. & Gallagher, E. 2016. The impact of education programmes on learning and school participation in low- and middle-income countries, International Initiative for Impact Evaluation (3ie), Systematic review summary 7. (Also available at http://www.3ieimpact.org/evidence-hub/publications/systematic-review-summaries/impact-education-programmes-learning-school-participation-low-and-middle-income-countries).

Sumberg, J. & Sabates-Wheeler, R. 2011. Linking agricultural development to school feeding in sub-Saharan Africa: theoretical perspectives. *Food Policy*, *36*(3), 341–349. https://doi.org/10.1021/cen-09507-scitech2.

Swensson, L.F.J. & Klug, I. 2017. *Implementation of decentralised food procurement programmes and the impact of the policy, institutional and legal enabling environment: the case of PRONAE and PAA Africa in Mozambique.* Brazil, International Policy Centre for Inclusive Growth (IPC-IG), No. 161).

Tranchant, J.-P., Gelli, A., Bliznashka, L., Diallo, A. S., Sacko, M., Assima, A., … Masset, E. 2018. The impact of food assistance on food insecure populations during conflict: Evidence from a quasi-experiment in Mali. *World Development*. https://doi.org/10.1016/j.worlddev.2018.01.027.

Upton, J.B., Lentz, E.C. & Barrett, C.B. 2012. *Local food for local schools : An analysis of the impact of local procurement for a school feeding programme in Burkina Faso* (Vol. 6893).

v**an Stuijvenberg, M., Dhansay, M., Smuts, C., Lombard, C., Jogessar, V. & Benadé, A.** 2001. Long-term evaluation of a micronutrient-fortified biscuit used for addressing micronutrient deficiencies in primary school children. *Public Health Nutrition*, *4*(06), 1201–1209. https://doi.org/10.1079/PHN2001179.

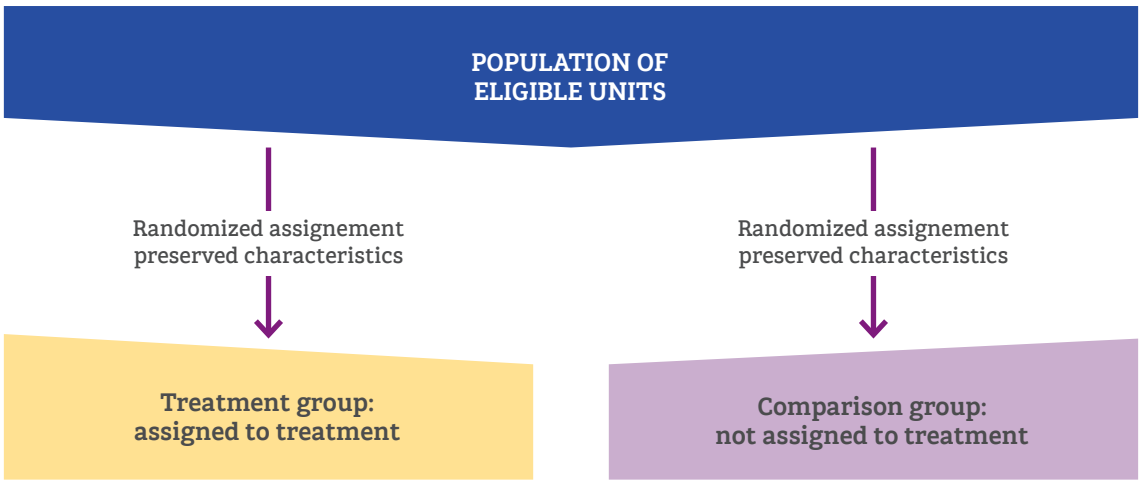
WFP. 2013. *State of School Feeding Worldwide 2013*. Rome, World Food Programme.

WFP. 2017. *Homegrown School Feeding. A framework to link school feeding with local agricultural production*. Rome, World Food Programme.

White, H. & Raitzer, D.A. 2017. *Impact evaluation of development interventions: A Practical Guide*. Asian Development Bank. https://doi.org/10.22617/TCS179188-2.

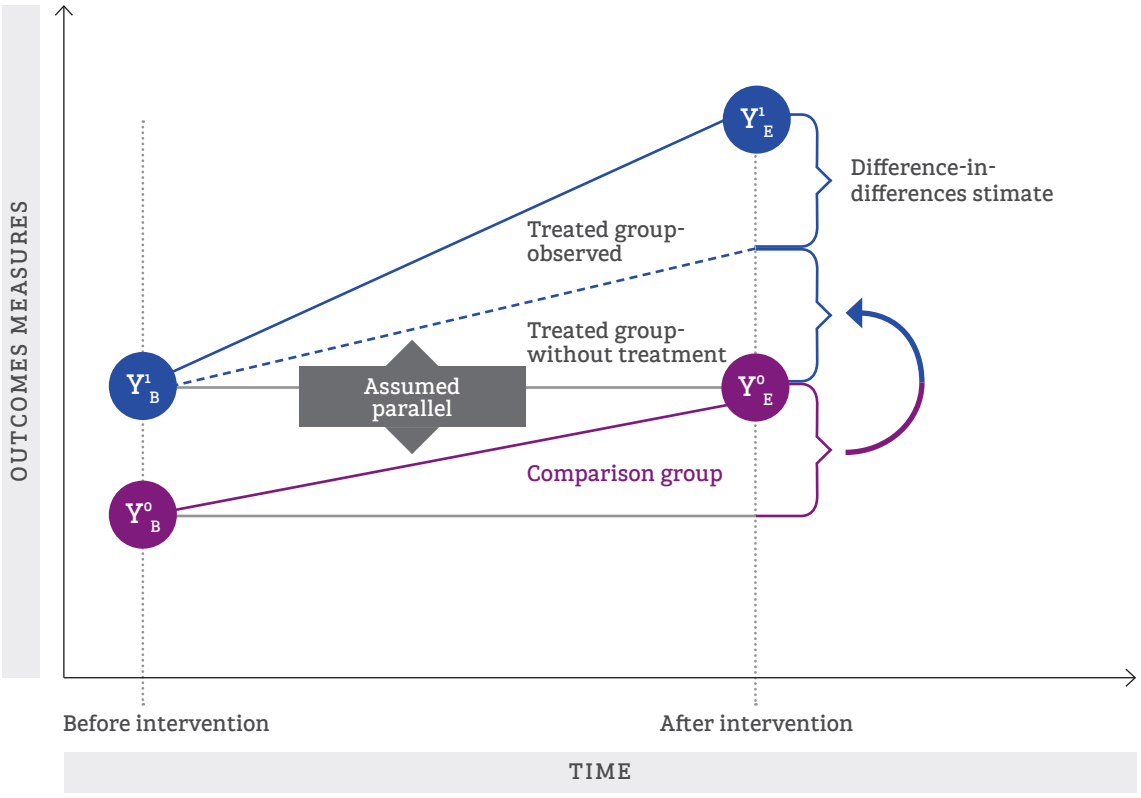
Appendix A
Randomization scheme

FIGURE 5. **CHARACTERISTICS OF GROUPS UNDER RANDOMIZED ASSIGNMENT**



Source: Gertler *et al.*, 2016.

FIGURE 6. **ILLUSTRATION OF DIFFERENCE-IN-DIFFERENCES**



Source: White and Raitzer, 2017.

Appendix B

Power calculation for determining sample size

SIMPLE DESIGN

As discussed in the main text, the minimum detectable effect depends upon the *t*–statistic values for the significance level α and the chosen level of power (1-β), the standard error of the outcome variable σ_y, the proportion of the sample in the treatment group (P), and the sample size (n):

$$\text{MDE}=(t_{\frac{\alpha}{2}}+ t_{1-\beta})\sigma_y\sqrt{\frac{1}{P(1-P)n}}$$

We can obtain sample size $n=\frac{(t_{\frac{\alpha}{2}}+ t_{1-\beta})^2\sigma_y^2}{MDE^2P(1-P)}$

In case a stratified sample design is adopted to assure representation of population subgroups in the sample, power analysis determining the sample size for each specific strata should be conducted separately. In case outcome variables and intervention effects are expected to vary across groups, the resulting sample size would be different across subgroups.

CLUSTER DESIGN

Cluster designs provides that the unit of assignment contain multiple units for which the data are collected. This has relevant implication for sample size. The intra-cluster correlation (ICC) coefficient ρ, is a measure of how similar the units are *within* each cluster. Power is higher the more heterogeneous the units are within a cluster, as reflected in a lower. The ICC is calculated as:

$$\rho = \frac{s_b^2}{s_b^2 + s_w^2}$$

where s_b^2 is the variance of the outcome variable between clusters, and s_w^2 is the variance of the outcome variable within clusters. The ICC is therefore the fraction of the total variance that is between clusters. When there is no interdependence between individuals within a cluster, the ICC is 0. Ideally, the best source for the ICC to use in power calculations is from a dataset similar to the one that will be used in the evaluation with the same outcome variable, the same type of cluster, and covering the same population. A second source is from previous research contributions or pre-analysis plans.

Sample size calculated ignoring statistical dependence within clusters, needs to be multiplied a *design effect DE*.

$$DE = 1 + (m\text{-}1)\ \rho$$

where m is the number of individuals per cluster and ρ is the ICC. Thus, the true sample size needed, accounting for intra-cluster correlation, is:

$$n = \frac{(t_{\frac{\alpha}{2}} + t_{1-\beta})^2 \sigma_y^2}{MDE^2 P(1\text{-}P)}\ 1 + (\text{m-}1)\ \rho$$

Two observations:

- cluster design requires more observations than a simple design;
- the number of clusters is the main factor determining the power of a study for a clustered intervention, rather than the number of observations in each cluster.

Similarly to what said for simple designs, in case of stratified cluster sampling power analysis determining the sample size for each specific strata should be conducted separately.

Home Grown School Feeding programs have seen a considerable growth around the world in recognition of their crucial role as boosters of children health and educational outcomes as well as a nation's overall future growth potential, while also stimulating current economic activity and developing markets through local procurements. The rigorous evaluation of the effects of these programs on children's and local economy's outcomes poses several challenges due to the presence of multiple treatment arms, complex targeting criteria and the difficulties from lack of treatment randomization. This work brings together the most up-to-date statistical techniques for program evaluation and the experience of FAO in setting up the research design, the data collection process and in analyzing the data for the purpose of evidence-based policy advocacy.

ISBN 978-92-5-135886-3

9 789251 358863

CB8970EN/1/03.22