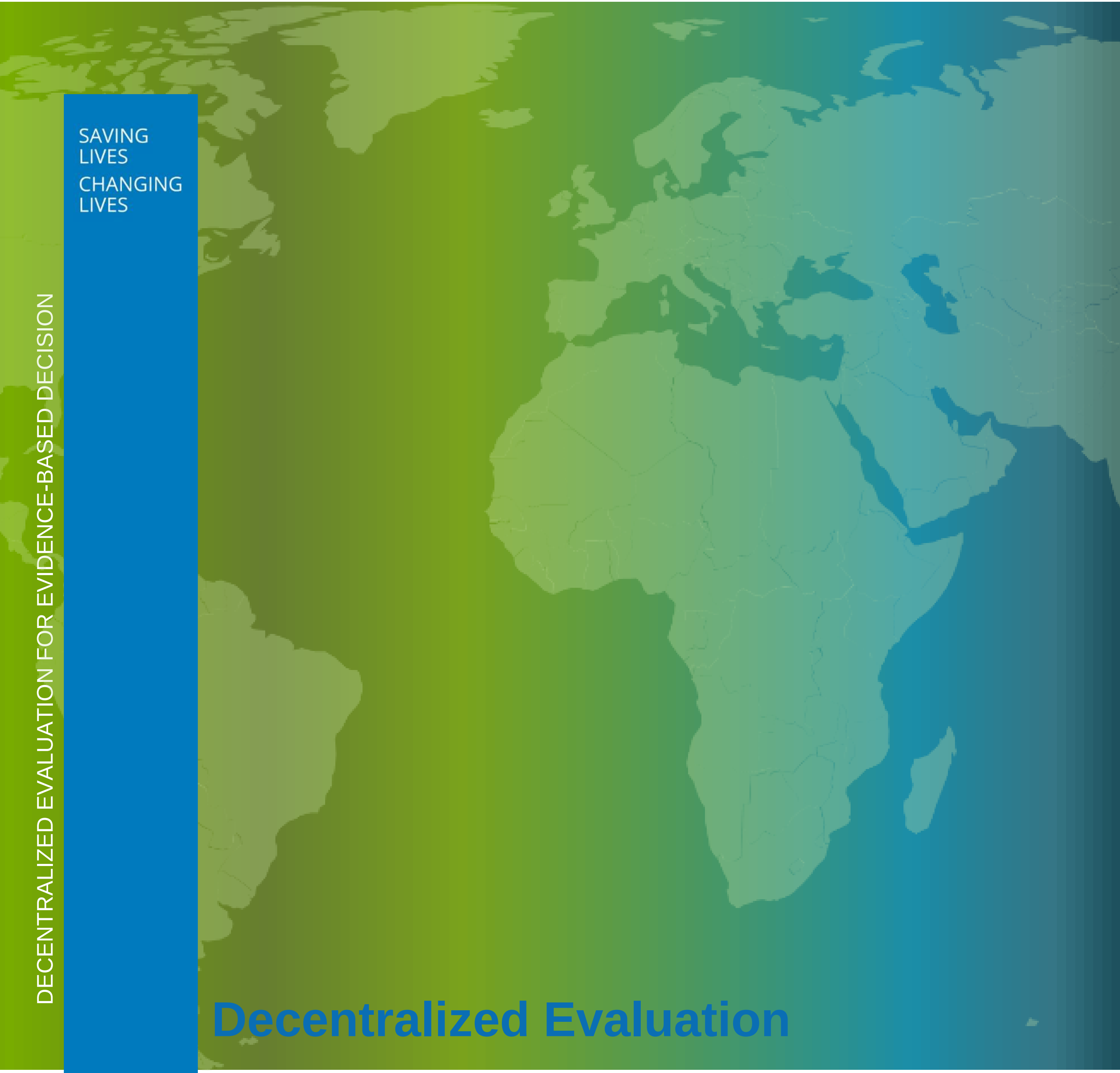


SAVING
LIVES
CHANGING
LIVES

DECENTRALIZED EVALUATION FOR EVIDENCE-BASED DECISION

Decentralized Evaluation

Impact Evaluation of WFP's Fresh Food Voucher Pilot Programme in Ethiopia

10/2017-1/2019

Evaluation Report

May 2019
WFP Ethiopia Country Office
Evaluation Manager: Roberto Borlini

Prepared by
Dr Kalle Hirvonen, Team Leader
Dr Kaleab Baye, Deputy Team Leader
Mrs Woinshet Tizazu Abate, Project Nutritionist

World Food Programme

Acknowledgements

This work has been funded by the World Food Programme. We are grateful to Roberto Borlini, Fragrance Manyala, Salome Daniel, Alemu Berhe, Maha Khan, Andualem Agize, Samir Wanmali, Praemeenah Poobalan, Filippo Dibari and Sibi Lawson-Marriott of WFP for their support throughout this project. This evaluation has benefitted from comments from Anna Henttinen (WFP), Jonas Heirman (WFP) and Paul J. Christian (The World Bank). We thank Dr Paulos Getachew, Ato Tilahun Bekele, the supervisors Ato Getachew Adane, Ato Tegegn Bogale, and Ato Tewodros Negussieas well as their team of 15 enumerators who played a critical role in ensuring that the fieldwork was undertaken to a high standard and it is a pleasure to acknowledge their contribution here. We are grateful for the administrative support provided by the International Food Policy Research Institute's Ethiopian Strategic Support Program and Addis Ababa University. Finally, none of this work would have been possible without the openness and hospitality of the 574 households, 147 fresh food traders and 108 caregivers who sat with us through the quantitative or qualitative interviews answering our questions about themselves. We thank them all sincerely.

Disclaimer

The opinions expressed in this report are those of the Evaluation Team, and do not necessarily reflect those of the World Food Programme. Responsibility for the opinions expressed in this report rests solely with the authors. Publication of this document does not imply endorsement by WFP of the opinions expressed.

The designation employed and the presentation of material in maps do no imply the expression of any opinion whatsoever on the part of WFP concerning the legal or constitutional status of any country, territory or sea area, or concerning the delimitation of frontiers.

Table of Contents

Table of Figures

Executive Summary

This Endline Report is for the impact evaluation of WFP’s Fresh Food Voucher (FFV) Pilot Programme in the Amhara region of Ethiopia.

The FFV Pilot Programme focused on households with pregnant and lactating women and children under 2 years of age (6 to 23 months). The beneficiary households received a voucher that can be used to purchase fresh foods (fruits, vegetables and certain animal source foods).

The main objectives of the evaluation were to assess and report on the performance and results of the FFV programme to help WFP present high-quality and credible evidence of actual impact to its donors. In addition, the purpose of the evaluation was to determine the reasons why certain results occurred (or not) to draw lessons, derive good practices and pointers for learning.

The primary objective of the pilot programme was to improve dietary diversity among children between 6 and 23 months of age and among pregnant and lactating women (PLW). The focus on these demographic groups was based on the now well-established theory on the importance of the 'first 1,000 days'. Dietary outcomes within the first 1000 days in Ethiopia, and particularly in Amhara are extremely poor and improving these outcomes is high on the national policy agenda. Against this background, the intervention had three planned outcomes:

i. Pregnant and lactating women and children 6-23m adopt a healthier, more diverse diet.

ii. Knowledge, attitude and practices regarding access and use of nutritious foods improve.

iii. Local food markets are able to respond to the increased demand by increasing the supply (so as to not to increase prices) and availability of different fresh foods.

The pilot programme experimented with two transfer values. For each transfer there was a 'family-size adjustment' to take into account the fact that larger households have higher food needs than smaller households. A household with up to two members received a monthly voucher of $12 or $21; households with members between 3 and 5 received $14 or $23; and households with 6 and more members received $17 or $26.

This evaluation was commissioned by WFP's Country Office in Ethiopia and covered the period from 10/2017 to 1/2019 with survey phases repeated at one-year interval in December 2017 and December 2018. The expected users of this report are the WFP as well as the Government of Ethiopia, the UN Country team and the donors of the FFV project; KfW (Germany) and the Government of Australia.

This impact evaluation could not conclusively answer the evaluation questions it set out to answer. This was because the program was not implemented as designed. Most importantly, most households did not receive transfers on a monthly basis and this meant that the impact evaluation strategy devised by the evaluation team could not be used to answer the four evaluation questions. The evaluation team was unaware of these implementation issues before the endline surveys began and therefore could not react to this, for example, by postponing the endline survey.

1.1. Methodology

The evaluation was designed to assess the impact of the WFP's Fresh Food Voucher (FFV) program against the following evaluation criteria: effectiveness and impact. The main evaluation questions, as indicated in the Terms of Reference, were:

Q1. What are the differential impacts of the programme on diet diversity for the different voucher values?

Q2. What are the main changes in knowledge, attitude and practices of the beneficiary households regarding access and use of nutritious foods?

Q3. Which transfer value is more cost-effective in delivering nutritional results?

Q4. What are the impacts of the project on the local markets of fresh foods?

In order to respond to these questions, the evaluation team conducted a randomized controlled trial (RCT), combined with quantitative and qualitative data collection in December 2017 before the program began and when it was still operational in December 2018. A major limitation of this evaluation is that the program was not implemented as designed and therefore, the impact evaluation could not be carried out.

1.2. Key Findings

The key findings of the evaluation team are summarised below, structured according to the main evaluation questions and indicating the type and strength of evidence supporting each finding.

Q1. What are the differential impacts of the programme on diet diversity for the different voucher values?

- Unfortunately, the lack of implementation fidelity does not allow the evaluation team to answer this question. Descriptive analysis of the voucher purchasing patterns suggest that the larger voucher somewhat increased the likelihood that the households purchased more expensive fruits, such as mangoes and oranges that are rich in Vitamin A and C. However, no notable differences were found with respect to animal source foods, possibly because of the limited supply of these products among the FFV traders.

Q2. What are the main changes in knowledge, attitude and practices of the beneficiary households regarding access and use of nutritious foods?

- The social behavior change communication (SBCC) activities were not yet rolled out at the time of the endline survey. Consequently, there is little change in knowledge and attitudes. However, in terms of practices, the descriptive and qualitative data provide some suggestive evidence that among beneficiary households, FFV has improved access of fresh foods, particularly fruits and vegetables. These changes were however mostly observed by an increase in the consumption of onion and potatoes – items belonging to food groups which were widely consumed already prior to the intervention. The primary reason for selecting these food items relates to their better storability.

Q3. Which transfer value is more cost-effective in delivering nutritional results?

- Unfortunately, the limited implementation fidelity does not allow the evaluation team to answer this question.

Q4. What are the impacts of the project on the local markets of fresh foods?

- The evidence gathered from the qualitative interviews suggest that the project has improved the availability fruits and vegetables in the markets. However, there are fears that the prices have also increased after the FFV was introduced in these markets.

Although not explicitly mentioned in the evaluation matrix of the ToR, the evaluation team assessed various gender issues with the data that were collected. The findings from these analyses can be summarized as follows. First, the differences in children's dietary outcomes did not vary markedly by child's sex. Second, the dietary diversity between male and female caregivers were nearly identical before the intervention began. Third, in more than 60 percent of the household, the male caregiver is directly involved and supported the mother to feed the child. Finally, analysis of the qualitative data reveal that there is little engagement by men in the market. Asking about decision-making regarding consumption within the households, we observed that in most cases, women are the one making those decisions.

1.3. Overall conclusions

In response to the first evaluation criterium of effectiveness, the evaluation team concludes that the evidence that the FFV pilot program achieved its objectives is inconclusive. Similarly, in response to the second evaluation criterium of impact, the evaluation team concludes that the evidence that the FFV pilot program had impact on the primary outcomes is inconclusive. However, the serious implementation challenges with respect to key interventions (dissemination of transfers and provision of SBCC) strongly suggest that the FFV pilot had only limited impact.

1.4. Recommendations

Based on the findings and conclusions of this evaluation, the recommendations of the evaluation team are outlined below. The target group for each recommendation is clearly identified. The recommendations are structured by priority.

Recommendation 1: The WFP needs to make sure that all eligible households receive transfers on a monthly basis and that the transferred amounts are both predictable and consistent across months.

Recommendation 2: The WFP needs to hire dedicated program coordinators, if possible, at the kebele level to provide on-the-spot support to beneficiaries and eligible households that are not receiving vouchers. A complaint feedback mechanism that complementing the hotline number should be set-up.

Recommendation 3: The WFP should reconsider excluding these onions and potatoes from the voucher scheme to maximize the dietary quality impacts.

Recommendation 4: The WFP needs to make sure that all beneficiary households receive SBCC. The gap between perceived nutrition knowledge and actual nutrition knowledge need to be taken into account in the SBCC strategy.

Recommendation 5: The WFP should consider providing specific training on how to cook and preserve fresh foods.

Recommendation 6: The WFP needs to closely monitor the intervention markets to ensure that the quality of the foods supplied by the FFV vouchers and their prices remain reasonable.

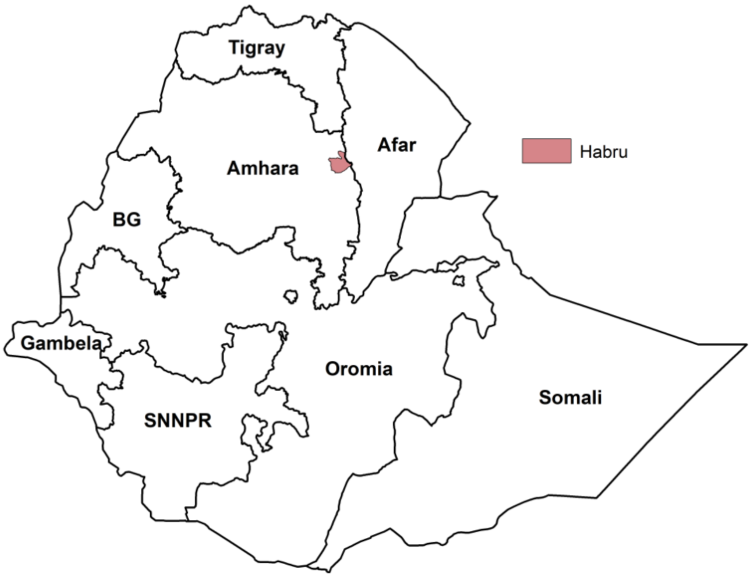
1. Introduction

1. This Endline Report is for the evaluation of WFP's Fresh Food Voucher (FFV) Pilot Programme in the Amhara region of Ethiopia. The FFV Pilot Programme focused on households with pregnant and lactating women and children under 2 years of age (6 to 23 months). The beneficiary households received a voucher that can be used to purchase fresh foods (fruits, vegetables and certain animal source foods). The FFV Pilot Programme was supposed to begin in January 2018 but due to external events (insecurity in the implementation area), the actual implementation began in June 2018. The pilot phase ended in January 2019 when the program expanded to other districts in the same region.
2. Due to limited implementation fidelity (described below), this impact evaluation could not conclusively answer the evaluation questions it set out to answer.
3. This evaluation was commissioned by WFP's Country Office in Ethiopia and covered the period from 10/2017 to 1/2019 with survey phases repeated at one-year interval in December 2017 and December 2018.
4. The main objectives of the evaluation were to assess and report on the performance and results of the FFV programme to help WFP present high-quality and credible evidence of actual impact to its donors. In addition, the purpose of the evaluation was to determine the reasons why certain results occurred (or not) to draw lessons, derive good practices and pointers for learning. This report provides evidence-based findings to inform operational and strategic decision-making. Findings will be actively disseminated, and lessons will be incorporated into relevant lesson sharing systems. Given the pilot character of the intervention, a strong emphasis was placed on the learning objective. The other key objective of the programme was accountability.
5. The expected users for this Endline Report are the WFP as well as the Government of Ethiopia, the UN Country team and the donors of the FFV project; KfW (Germany) and the Government of Australia.

1.1. Overview of the Evaluation Subject

6. WFP's Fresh Food Voucher (FFV) Pilot Programme took place in 12 Productive Safety Net Program (PSNP) kebeles (sub-district) in the Habru woreda (district) in the Amhara region of Ethiopia; see Figure 1. At the time of the last census in 2007, the total population of Habru was 192,742 (CSA 2010). More than 88 percent of the population (171,142) resided in a rural area. In terms of religion, 76.8 percent of the population is Muslim while 23.0 percent are Orthodox Christian.

Figure 1: Location of Habru



Source: Authors' construction. BG = Benishangul Gumuz.

7. The FFV targets pregnant and lactating women and households with children under 2 years of age (6 to 23 months). The program beneficiaries receive a voucher that can be used to purchase fruits, vegetables and certain animal source foods (such as eggs and milk). To ensure that the markets are able to respond to the increased demand, the FFV pilot actively engages with fresh food traders in the intervention areas. More specifically, the FFV programme implements a retail engagement strategy to ensure that traders have the capacity to scale up their traded volumes and minimize food losses, and to make sure that local fresh foods are available throughout the year.
8. The transfer value was determined by the WFP and was based on a market assessment that involved interviews with households, traders and government officials in the pilot woredas. This pilot study experimented with two transfer values. For each transfer there was a 'family-size adjustment' to take into account the fact that larger households have higher food needs than smaller households. A household with up to two members received a monthly voucher of $12 or $21; households with members between 3 and 5 received $14 or $23; and households with 6 and more members received $17 or $26 (Table 1). The larger value vouchers were expected to permit households to afford animal source foods (ASFs) on top of fresh fruits and vegetables.

Table 1: Fresh Food Voucher values by household size

Household size	Value USD for Group 1	Value USD for Group 2
Up to 2	12	21
3 to 5	14	23
6 and above	17	26

9. WFP partnered with Hello-Cash, a local organization specializing in mobile banking, to deliver the vouchers. The beneficiary households received a text to their mobile phone when a voucher had arrived to their account. After this, the beneficiaries could go to the market and purchase fresh foods with their electronic voucher from retailers that were part of the FFV program. Within households, the targeted voucher recipient was the pregnant woman or mother of a young child. In the Ethiopian context, this person is typically the spouse of the household head, although sometimes she is the head herself if the father/husband is absent.

10. The FFV pilot was also planned to include a sizable behavioral change component that attempts to address the knowledge constraints related to nutritious diets. Within this component, the beneficiaries were expected to regularly attend nutrition education sessions to receive information about the importance of a diverse diet, along with other social behavior change communication (SBCC) initiatives such as practical demonstrations on the proper usage of nutritious foods. However, the implementation of the SBCC activities were severely delayed and were just rolled out at the time of the endline surveys. The planned SBCC strategy is described in more detail in Annex 1.

11. The primary objective of the pilot programme was to improve dietary diversity among children between 6 and 23 months of age and among pregnant and lactating women (PLW). The focus on these demographic groups is based on the now well-established theory on the importance of the 'first 1,000 days'. This period, spanning the pregnancy and the first 2 years of child's life, is considered critical for child's physical growth and cognitive development. As such, this period offers a critical window to shape long-term health and nutrition outcomes – many of which are considered difficult to reverse later in life.

12. Against this background, the intervention has three planned outcomes:

i. Pregnant and lactating women and children 6-23m adopt a healthier, more diverse diet.

ii. Knowledge, attitude and practices regarding access and use of nutritious foods improve.

iii. Local food markets are able to respond to the increased demand by increasing the supply (so as to not to increase prices) and availability of different fresh foods.

13. The total budget for the pilot programme is 4.3 million USD, funded by KfW (Germany) and the Government of Australia. It is understood that at the time of the endline survey, a significant fraction of this budget was not yet spent.

14. The total number of households with a pregnant or lactating woman or children 6-23 months of age reached is 11,000. The initial plan was that the programme first focuses on up to 5,000 households in the Habru district and within four months scale up to the full 11,000 households in the other two districts. However, this geographical expansion of the pilot program was delayed and only began in 2019. Of note is that due to budget constraints, the total number reached by the project is less than the estimated number of eligible households in the three districts. In this context, the WFP together with the evaluation team randomly allocated villages into three groups: those receiving smaller-sized voucher; those receiving larger-sized voucher and those receiving no vouchers. Outside Habru, where the randomized control trial took place, the WFP decided which villages or kebeles were part of the program. This decision was largely driven by assessment of needs within the local population, functionality of the markets and implementation costs.

15. The logic of the programme is provided in Annex 2 (Theory of Change). The main outcome of interest in this evaluation was the quality of diets adopted by PLW and children less than two years of age (green box in the ToC). The other evaluation questions focus on changes in KAP (yellow boxes) and impacts on markets (purple boxes). In addition, the evaluation design also considered the cost effectiveness by considering how the monetary size of the voucher affects food consumption patterns.

1.2. Context

16. With Gross National Income of just below $600, Ethiopia is one of the poorest countries in the world. In 2011, 31 percent of the population fell below the $1.25 poverty line (World Bank 2015) and recent FAO et al. (2017) estimates suggest that 29 percent of the Ethiopian population are undernourished. Moreover, large part of the population depends on rain fed agriculture or pastoralism rendering livelihoods vulnerable to droughts. The most recent large-scale drought occurred in 2015 and it was estimated that more than 10 million people were in need of food aid (GoE and Ethiopia Humanitarian Country Team 2016). Recent years have seen rise in inter-communal conflicts that have negatively affected food security in the country (NDRMC 2018), including the Amhara region where the FFV program operates. It is widely believed that the implementation of various development and humanitarian programs have become more challenging in some areas of the country because of these inter-communal conflicts.

17. Despite considerable improvements over the last decade, chronic under-nutrition rates remain high in Ethiopia. In 2016, more than 38 percent of children under age 5 are

chronically under-nourished (short for their age; stunted) and 22 percent of women in reproductive age are underweight (body-mass index below 18.5) (CSA and ICF 2016).

18. Low dietary diversity is considered a risk factor for chronic under-nutrition, micro-nutrient deficiencies and non-communicable diseases. In Ethiopia, only 14 percent of children 6-23 months of age have an adequately diverse diet (4 or more food groups out of 7) and only 7 percent of children in the same age range meet the WHO criteria for a minimum acceptable diet (MAD) (CSA and ICF 2016). Perhaps indicative of low dietary diversity, 57 percent of Ethiopian children aged 6-59 months had hemoglobin levels below 11 g/dl (CSA and ICF 2016) – a marker for anemia. Moreover, recent survey conducted by IFPRI in food insecure areas in the Ethiopian highlands showed that only 1.8 percent of women in reproductive age meet the minimum dietary diversity for women (MDD-W) of 5 food groups or more out of 10 (Berhane et al. 2017).

19. WFP supports the Government of Ethiopia in achieving SDG2 (achieving zero hunger) and SDG17 (partnership to support implementation of the SDGs). Achieving these ambitious goals require a high level of technical expertise and operational capability. WFP has a comparative advantage in providing such support to the Ethiopian Government as illustrated by its range of life-saving and resilience-building interventions that use food, cash, nutrition assistance, and innovative market-based approaches. For example, WFP currently assists about 600,000 refugees from neighboring countries, provides nutrition assistance to 1.6 million vulnerable people in emergency settings, and supports the Ethiopia's Productive Safety Net Programme (PSNP) that provides food and cash to vulnerable people during the lean season in exchange of public works (e.g. natural resource management, road construction, etc.).

20. WFP is gradually moving away from emergency-focused/project-based to a more holistic portfolio approach that aims to prevent chronic undernutrition through interventions that address the underlying causes of food and nutrition insecurity. First, it is building the Government's proactive disaster risk management through WFPs early warning tools developed by the Vulnerability Assessment and Mapping (VAM) unit; second, it is breaking the artificial walls between "humanitarian" and "development" work. An illustration is the Purchase for Progress programme, where WFP buys food for humanitarian relief from local smallholder farmers making the interventions not only cost-efficient, but also boosting the local economy, strengthening small holder farmers', and supporting local markets. In addition, WFP's efforts to improve humanitarian relief delivery in close collaboration with the Ethiopian Government (e.g. Ethiopian Maritime Affairs Authority and The Federal Road Transport Authority) has strengthened the logistics capacity of the country.

21. In its effort to improve poor diets in Ethiopia, WFP’s school feeding programme in partnership with the Government has been providing nutritious meals for about 400,000 school children. This will prevent chronic undernutrition among school children, but because stunting reaches its peak during the complementary feeding period (6-23 months age) much of the damage has already happened by then. Several interventions that aim to improve complementary diets have been implemented by various actors, but these have not lead to substantial and sustainable improvements. This calls for innovative and multi-dimensional approaches to address this multi-faceted and complex challenge. Particularly, understanding the role of income, strengthening local markets and SBCC in supporting nutritious diets is critical.
22. The FFV intervention takes place in the Amhara region of Ethiopia. Amhara provides a good location for this pilot study. The region is characterized by extremely low levels of dietary diversity. Only 1.3 percent of the children meet MAD and 2.7 percent the criteria for minimum dietary diversity. Out of the 11 administrative regions in Ethiopia, these dietary scores are among the lowest in the country. Moreover, recent large-scale IFPRI survey in food-insecure districts showed that none of the women of reproductive age in Amhara met MDD-W (Berhane et al. 2017).
23. Recent research from Ethiopia suggests that low dietary diversity in Ethiopia is driven by a combination of different factors. First, a number of studies now show that lack of knowledge of the health benefits associated with diverse diets (Abebe, Haki, and Baye 2016; Kim et al. 2016; Zerfu, Umeta, and Baye 2016) is poor. Second, another branch of research identifies poor access and availability to nutritious foods as an important constraint (Abay and Hirvonen 2017; Stifel and Minten 2017; Headey et al. 2018). Finally, a recent IFPRI study (Bachewe et al. 2017) showed that the real prices of nutrient-dense food groups such as fruits, vegetables and animal source foods increased between 19-62% in just a decade (2007-2016). These price trends put forward a third hypothesis: many poor households simply cannot afford to consume a nutritionally rich and diverse diet – a finding supported by recent national level analysis by Hirvonen, Wolle, and Minten (2018).
24. The design of the WFP's FFV pilot programme aims to address these constraints. Focusing on young children and pregnant/lactating women, the programme relaxes households' budget constraints through the provision of vouchers that permit households to purchase fruits, vegetables and animal source foods. The program also has a behavioral change component to address the knowledge constraints. Moreover, WFP actively works with the food traders to ensure that the markets are able to respond to the increased demand of fruits, vegetables and animal source foods.

25. The programme complements various ongoing efforts in the country to address the root causes of high under-nutrition, food security and low dietary diversity as well as to stimulate economic growth. Most notably, the FFV complements Ethiopia's flagship safety net program –PSNP – that reaches more than 7 million food insecure people in eight regions of the country, including Amhara. The beneficiary households receive cash or food payments against public works that take place during the slack season while households with limited labor capacity receive direct support. The payments are equivalent to 15kg of cereal and 4kg of pulses per person per month.[1] Earlier IFPRI-led evaluations show that the PSNP has improved household-level food security and helped households to protect their assets (Berhane et al. 2014; Berhane, Hirvonen, and Hoddinott 2016) – but has not improved child health and nutrition outcomes (Berhane, Hoddinott, et al. 2016). To this end, the current phase[2] of the PSNP has been designed to be nutrition-sensitive (GFDRE 2014a, 2014b), by linking nutrition behavioral change communication sessions to the public works component of the programme. The PSNP is supported by the WFP. In the PSNP, the transfers are made either in food or cash. The food component in the PSNP focuses on cereals, which provide calories but contain micronutrient in low densities. In contrast, the FFV program places a considerable focus on micro-nutrients by restricting the transfers to the purchase of fruits and vegetables.

26. The FFV program is also well-aligned with the key nutrition policies in the country. First, the National Nutrition Programme sets a target to increase the proportion of children 6-23 months that meet MAD to 40 percent by 2020 (GFDRE 2016). Second, the government recently drafted its first ever Agriculture Sensitive Nutrition Strategy to further tackle high under-nutrition and low dietary diversity. Akin to the WFP's FFV pilot programme, the Agriculture Sensitive Nutrition Strategy places an emphasis on market development, demand creation and nutrition education to reach these objectives (MoANR and MoLF 2016). Third, one of the core strategic objectives in agricultural sector within the new Growth and Transformation Plan is to improve agricultural marketing systems in the country (FDRE 2017).

[1] It is understood that the pulses are no longer part of the food transfers.

[2] The PSNP was initially planned as a three-year intervention, from 2005 to 2007. At the end of this 1st phase, the program was reviewed by both government and donors and based on this review, a 2nd phase was implemented between 2008 and 2010. Subsequently, this process of assessment and renewal has continued with the third phase (2011 to 2015). The fourth phase scheduled to operate from 2016 to 2020. Each phase added some new elements to the program. For example, much work was done to improve the targeting of the programme. During phase 3, the PSNP expanded to the pastoralist areas of the country (Afar and Somali regions) and paid more attention to beneficiary households' graduation. Phase 4 made the program nutrition-sensitive.

27. Apart from the Government of Ethiopia, the key partners are Ethiopia's development partners and non-government organizations working on food security and nutrition, including UNICEF, the World Bank, DfID, USAID, Save the Children, Global Alliance for Improved Nutrition (GAIN) and Scaling Up Nutrition (SUN).

28. Finally, with the aim of improving the dietary quality of pregnant and lactating women, the intervention also addresses gender inequality in the country. Out of 188 countries, Ethiopia ranks 116 in UNDP's gender inequality index (UNDP 2016). Maternal mortality rates remain high with 353 deaths per 100,000 live births and only 11% of women (21% men) aged 25 or more have attended secondary school (UNDP 2016). In addition, earlier research from rural Ethiopia shows that, in poor households, adult women's food consumption is often first sacrificed when households' face economic shocks (Dercon and Krishnan 2000). The extent to which still occurs in rural areas is unknown but by subsidizing household food consumption, the FFV programme has also the potential to address this issue.

1.3. Evaluation Methodology and Limitations

Description of the methodology

29. The evaluation team was requested to answer four questions listed in the TOR:

Q1. What are the differential impacts of the programme on diet diversity for the different voucher values?

Q2. What are the main changes in knowledge, attitude and practices of the beneficiary households regarding access and use of nutritious foods?

Q3. Which transfer value is more cost-effective in delivering nutritional results?

Q4. What are the impacts of the project on the local markets of fresh foods?

30. It was agreed with the WFP that the evaluation focuses on eligible households with children 6-17 months of age at the baseline, and their mothers (most of which were still lactating). The TOR further defined the main outcomes of interest as minimum acceptable diet score for children 6-23 months (MAD) and minimum diet diversity for women of reproductive age (MDD-W). We used the best practices devised by the WHO, FAO, and FHI-360 to collect and construct these outcome variables (see WHO 2008; FAO and FHI 360 2016).

31. The primary goal of the evaluation was to measure the *causal* impact of the WFP's Fresh Food Voucher Pilot Programme as differences in outcomes between the program beneficiaries (i.e. those that receive vouchers) and their counterfactual, a measure of what outcomes would have been for this group had they not received the program. Most evaluation strategies are designed to find a method for constructing a proxy for these

counterfactual outcomes from information on non-beneficiaries. This requires controlling for the effects of confounding economic and other contextual factors that make program beneficiaries systematically different from an average non-beneficiary. These confounding factors can include the relative poverty of program beneficiaries in targeted programs, exposure to economic shocks, or differences in household characteristics (e.g. demographics, skill levels, or social networks) that affect the impacts of the program. Typically, addressing the confounding factors requires a control group. In this context, this means a group of households that do not receive vouchers, only SBCC. This is feasible given that the total number of eligible households in the three woredas is considerably larger than the number of intended beneficiaries of the pilot program (11,000 HHs).

32. Within this framework, the evaluation team conducted a Cluster Randomized Controlled Trial (RCT) with two treatment arms and one control arm. The reliability and internal validity of the findings based on the RCT approach is considered high (Duflo, Glennerster, and Kremer 2007)

33. A cluster in this evaluation was defined as a village (*got* in Amharic) and these clusters were randomly allocated to the different arms of the trial. Within a cluster, all eligible households received the same treatment.

34. As described above, the 1st treatment arm received vouchers worth of $12 to $17 (depending on the household size) and the 2nd treatment arm received vouchers worth of $21 to $26. Control households were not supposed to receive vouchers. Importantly, all three arms were expected to receive SBCC, for example, cooking demonstrations on the proper usage of nutritious foods. However, as we explain below in this section, the SBCC activities were not yet widely rolled out at the time of the endline survey.

35. The RCT approach randomly allocated eligible households into treatment and control arms. Such random assignment overcomes the problem of constructing a valid counterfactual; the three groups are – in expectation – identical before the programme begins. The randomization was done in four steps using Stata (version 15). First, we created a list of all eligible villages[3] in Habru district. Second, we gave a random number to each village. Third, we sorted the data set using this randomly generated number from the smallest to the largest. Fourth, we picked the first 20 villages to control, the next 20 villages to treatment #1 (small voucher) and the following 20 villages to treatment #2 (large voucher).

[3] That is all villages located in kebeles (sub-districts) in which PSNP was operational.

36. The quantitative impact evaluation collected data from the three groups at the household level in two rounds. Quantitative data were collected in December 2017 before the intervention began (the “baseline” or “before” data), and again in December 2018 (the “endline” or “after” data). This survey setup permits a “before/after” comparison. These data were collected from households receiving vouchers and those that do not receive programs (“with the program” / “without the program”). By collecting data before and after the intervention on households with and without the program, it is possible to control for differences in baseline characteristics between beneficiaries and non-beneficiaries.

37. With data collected before and after the start of the intervention on households with and without the program, it is possible to estimate impacts of the program FFV using various well-known treatment effect models, such as “difference-in-differences” (DID). DID models estimate impacts as the difference in the change in outcomes between beneficiaries and non-beneficiaries.

38. The impact estimates then permit us to assess the cost-effectiveness of different voucher values by comparing the estimated impact to monetary value of the voucher.

39. Qualitative data were collected through Focus Group Discussions (FGDs) with caregivers and in-depth interviews traders. The interviews were conducted by trained and experienced interviewers using semi-structured questions. The objective of this qualitative survey is to *qualitatively* evaluate changes in a) caregiver's Knowledge, Attitude and Practices with respect to fresh foods and b) markets (interviews with traders) as a result of the FFV program. The overall study design of the qualitative assessment was based on the principles of qualitative research.

40. Although not explicitly mentioned in the evaluation matrix of the ToR, the evaluation study did include a gender dimension. First, the data on child level outcome variables – MDD, MMF and MAD – are reported separately by boys and girls during in the baseline and endline reports. Second, using the quantitative survey data, the evaluation team assessed gender dynamics by comparing diets between male and female caregivers in the households. Finally, we also assessed the role of the men in the decision-making processes with respect to child feeding.

Description of the sampling strategy

41. A detailed initial sampling strategy was provided in the Inception report. This strategy was revised afterwards to account for the revision in the age range of children from 6-23 months to 6-17 months and the fact that the baseline survey because the evaluation team was not able to find a sufficient number of children from the villages.

42. The minimum detectable effect sizes were set as follows:

(i) Non-voucher group (C): by endline 10% of the children meet MAD

(ii) Small voucher group (T1): by endline 25% of the children meet MAD

(iii) Large voucher group (T2): by endline 40% of the children meet MAD

43. We set significance level (α) at 0.05, power at 0.8 and accounted for intra-cluster correlation (0.03). The total number of clusters (villages) was set to 60. The calculations were done using Stata 15.0 and user-written 'clustersampsi' command.

44. Three study arms give rise to three different tests:

a) Non-voucher vs Small voucher

b) Non-voucher vs Large voucher

c) Small voucher vs Large voucher

45. These three tests require different sample sizes. The third test involving the two treatment groups requires the largest sample while the two tests involving the control (non-voucher) group require considerably smaller sample sizes. We therefore set different sample sizes for control and for treatment groups. The required sample size (after accounting for 10% attrition) for the non-voucher group was calculated as 140 households and for treatment households 220 in each arm; 580 in total for both groups. Totaling these numbers means that the total sample needed for this evaluation is 615 households. The breakdown of the sample size by study arm based on the original statistical power calculations is provided in Table 2. The actual, achieved sample sizes are discussed in Annex 3 along with other survey outcomes. In Annex 4 we show that the outcome variables and key households characteristics were balanced at the baseline, before the intervention began.

Table 2: Required sample sizes, adjusted for attrition

Arm	N	Clusters	N per cluster
Non-voucher	140	20	7
Small voucher	220	20	11
Large voucher	220	20	11
Total	**580**	**60**	**n/a**

Note: N refers to sample size. n/a = not applicable.

46. Finally, we note that the villages or the group of villages where the sample are drawn are of different sizes. There are two ways of accounting for this. The first method is to do a 'probability-proportional-to-size' sampling in which households in larger clusters have a higher probability of being selected, and *vice versa*. The second method is to use sampling weights when the data are analyzed. These weights adjust for the differences in sampling probabilities across clusters, by basically assigning more weight for observations originating from larger clusters. We have the size of the cluster (i.e.

number of households with young children) from the listing phase. We therefore apply this weighing strategy when we report our findings.

Data collection methods and tools

The evaluation used mixed (quantitative and qualitative) methods. The quantitative and qualitative data have been analyzed in parallel, with continuous dialogue over preliminary findings and complementary analysis to triangulate qualitative findings with quantitative analysis, and vice versa. The quantitative data were collected using paper questionnaires and these data were then carefully entered into an electronic format. After that the data were carefully cleaned and any inconsistencies solved by the evaluation team before analysis.

47. Annex 5 provides the Evaluation Matrix that maps these four research questions to the outcome variables, survey instruments and analytical methods. Below, we provide a brief description of how we planned to answer each evaluation question:

Q1. What are the differential impacts of the programme on diet diversity for the different voucher values?

- To answer this research question, we planned to use the baseline and endline quantitative data. With data collected before and after the start of the intervention on households with and without the program, we can estimate impacts of the program using difference-in-differences (DID) treatment model.

Q2. What are the main changes in knowledge, attitude and practices (KAP) of the beneficiary households regarding access and use of nutritious foods?

- The SBCC treatment does not vary across study arms and therefore, we cannot use the DID model to estimate impact. Given this, we planned to simply compare KAP outcomes before and after using our quantitative data. This approach does not allow to assess the causal impact of the FFV program on these outcomes. We supplement the analysis with qualitative data collected through FGDs with the caregivers during the baseline and endline.

Q3. Which transfer value is more cost-effective in delivering nutritional results?

- Using the impact estimates obtained after answering question 1, we can compare the voucher values to the achieved impact of different vouchers values to estimate cost-effectiveness of the program.

Q4. What are the impacts of the project on the local markets of fresh foods?

- To answer this question, we would need to compare the outcomes in FFV intervention markets to control markets. Ideally, the markets would have been randomly assigned into treatment and control. With only four intervention markets and without control markets, we cannot provide a causal estimate on the impact of FFV on local food markets. We therefore use the qualitative evidence gathered through FGDs with the traders and caregivers to qualitatively assess this issue.

Limitations

48. A major limitation of this evaluation is that the program was not implemented as planned and therefore the evaluation team could not carry out the impact evaluation as planned. Below we describe this issue in detail. We then describe the other limitations of the evaluation.

Assessing implementation fidelity

49. The evaluation strategy described above was devised with the assumption that the FFV program is implemented as designed. In the impact evaluation literature, implementation fidelity is defined as the degree to which the implementation of the program adheres to the original design (Breitenstein et al. 2010). Lack of implementation fidelity is typically the main reason why even theoretically sound programs fail to show positive impacts (Durlak and DuPre 2008; Kim et al. 2015).
50. Against this background, we use the endline data to assess the implementation fidelity of the FFV program. We begin by reminding that the theory of change on the FFV program (see Annex 2) relies on two interventions: SBCC activities and the dissemination of the fresh food vouchers. We examine the implementation of these two activities in turn.

Implementation of the SBCC activities

51. Our quantitative and qualitative endline data show that the SBCC activities were not yet widely rolled out at the time of the endline survey. This was also confirmed to us by the WFP.
52. The quantitative data reported in Table 3 shows that the exposure to SBCC is sporadic[4], and as a result, we see only a marginal improvement in caregivers' nutrition knowledge. The limited access to SBCC is also confirmed by the qualitative data. Caregivers reported to have received guidance to buy potatoes, onions, and tomatoes supposedly from the FFV program, partly because these foods are less perishable. Apart from the nutrition messages they are used to receive through the health extension program, the caregivers did not report any FFV specific nutrition education. Traders in Dire Roka

[4] The health extension workers are tasked to provide health and nutrition SBCC. The FFV project was supposed to strengthen the existing activities with specifically tailored messages around the use of fresh foods.

reported that households do not know how to process some of the vegetables. The area is arid and thus fresh foods were rarely consumed before the introduction of FFVs.

Table 3: Exposure to different SBCC sources

SBCC source (%)	baseline	endline
Caregiver had a contact with health extension worker in last 3 months	31.9	14.2
Caregiver attended a community event on nutrition in last 6 months	5.1	10.3
There was a cooking demonstration in the village in last 6 months	1.0	5.6
Received nutrition information from:		
Newspaper/magazine	1.9	2.1
Radio	6.3	5.2
TV	4.0	3.6
Poster/ banner/ board	0.9	0.6
Local theatre	0.5	1.7
Local loudspeaker	0.2	0.6
During a coffee ceremony	0.9	2.2
Mobile phone	1.2	1.7

53. The evaluation team was aware of the delay in the SBCC activities before the endline surveys were fielded. The WFP and the evaluation team discussed the option of delaying the endline survey because of this. However, it was agreed to go ahead with the original schedule because:

a) Children 'age out of the treatment eligibility'; the vouchers are targeted to households with children less than 2 years of age;

b) Many of the dietary indicators that are used as primary outcomes have not been validated for older children;

c) Previous work in this area suggests that SBCC works when implemented intensively over a relatively long period of time so this would have required to delay the survey by 3-6 months and this would have been problematic for a) and b);

d) Delaying the survey would have also meant that the baseline and the endline surveys would have occurred in different seasons, though this would have of course affected both treatment and control households

e) The health extension workers operating in the communities are expected to provide some SBCC as a part of their duties, though the implementation is often sporadic.

Implementation of the fresh food vouchers

54. The payment cycle data provided by the WFP shows that the first vouchers were only dispatched on 26 June 2018. As the FFV program was supposed to start in January 2018, this indicates a severe delay in implementation activities, possibly because of the insecurity in the Habru district during the first quarter of the year.

55. After the first voucher dispatchment, the payment cycle followed a bi-monthly (instead of the planned monthly) schedule: the 2nd cycle took place in 29 August, the 3rd cycle in 23 October and the 4th cycle in 10 December.

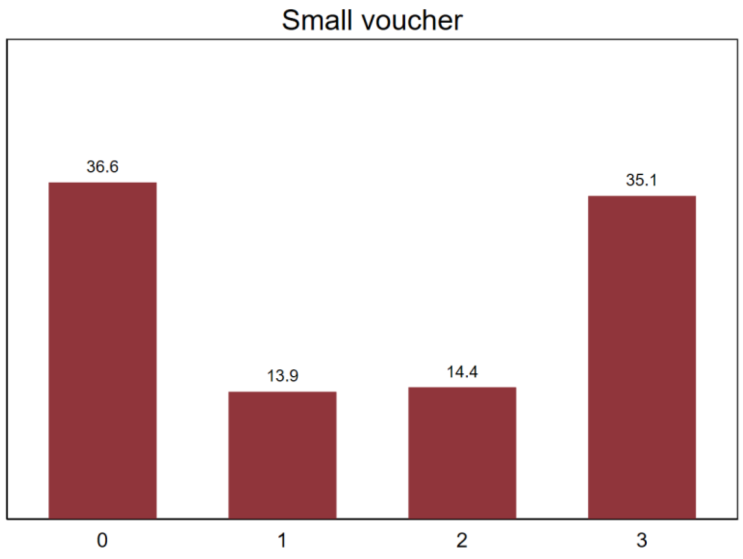
56. This irregular transfer schedule also becomes apparent in our quantitative data.[5] We also see that many of the households in the voucher groups never received a transfer. Table 4 shows that 38 percent of the households in the small voucher arm and 29 percent of the households in the large voucher arm did not receive any vouchers between June and December. Meanwhile, 16 percent of the households in the non-voucher group received a voucher. If we consider the month preceding the endline survey, we see that only about 11 percent of the households in the voucher arms received a voucher while less than 1 percent of the households in the control arm received a voucher.

Table 4: Percent of households receiving vouchers, by study arm

	Control	Small voucher	Large voucher
At least one voucher since July	15.6	62.4	70.7
At least one voucher since August	14.8	61.4	69.7
At least one voucher since September	14.8	59.9	67.7
At least one voucher since October	13.3	54.5	62.1
At least one voucher since November	0.7	10.4	10.6
At least one voucher since December	0.0	0.0	0.0

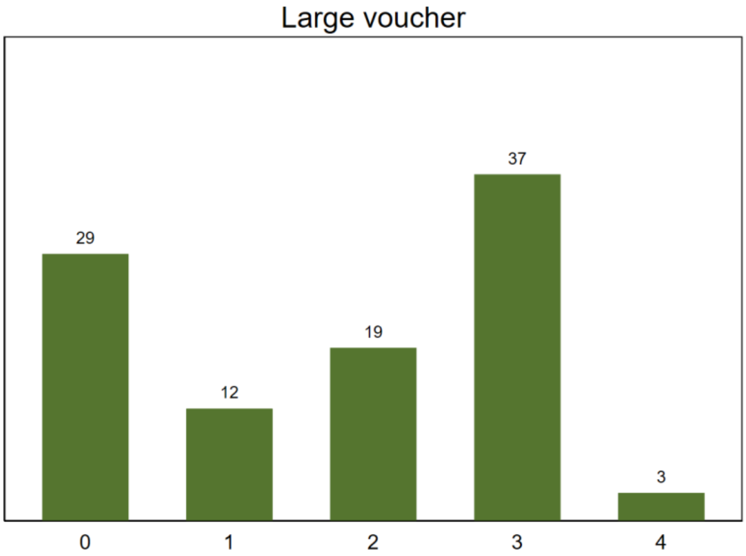
57. Figures 2-3 show the distribution in terms of vouchers received since June by household in small and large voucher arms, respectively. In both arms, about 35-40 percent of the households received 3 or more vouchers between June and December.

Figure 2: Total number of vouchers received, 'Small voucher' arm



[5] It is worth reminding here that the endline surveys took place between 5 and 20 December.

Figure 3: Total number of vouchers received, 'Large voucher' arm



58. The month-by-month examination shows that only 10-11 percent of the households in the two voucher arms received the transfer in November (Figures 4-5). A higher number of households received the transfers in the month of October (50 % for the lower and 60 % for the higher voucher). Meanwhile, about 13 percent of the non-voucher (control) households reported to have received a voucher in October (Figure 6).

59. The caregivers selected for the FGDs highlighted similar implementation challenges with respect to the payments. While the FFV was positively perceived by many caregivers, the inconsistency of the transfers was a source of complaint.

Figure 4: Percent of households receiving vouchers, Small voucher arm

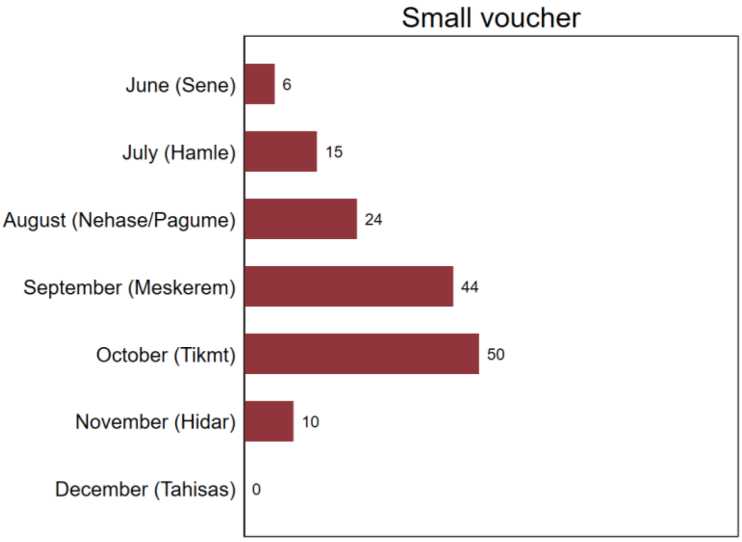


Figure 5: Percent of households receiving vouchers, Large voucher arm

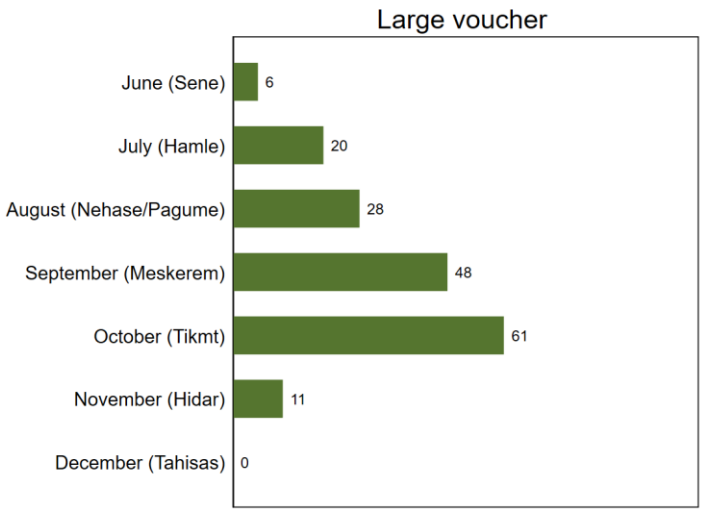
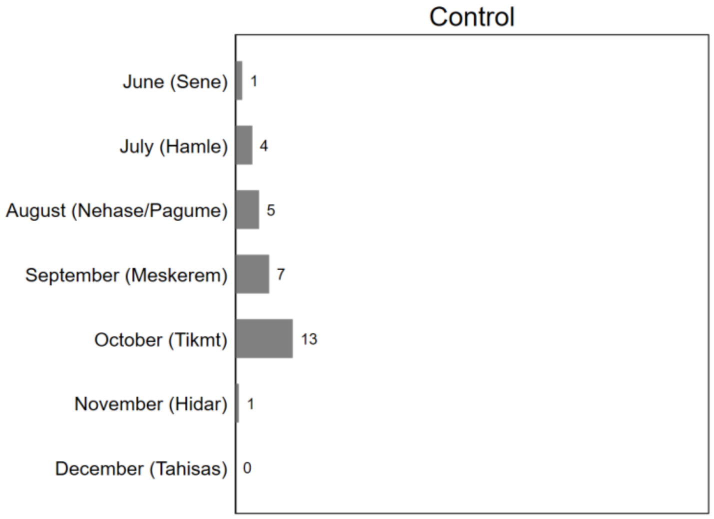


Figure 6: Percent of households receiving vouchers, Control arm



60. Before the fieldwork began for the endline, the evaluation team did communicate actively with the WFP about the transfers. From this communication, the evaluation team learned that some households received more than their entitlement in October. However, the evaluation team was unaware of the bi-monthly transfer schedule and the considerable under-coverage of the program. Because of the lack of information, the evaluation team could not react to this, for example, by postponing the endline survey.

Implications to the feasibility of the impact evaluation

61. At the core of the Program's Theory of Change (presented in Annex 2) is the assumption that improved access to fresh foods (through vouchers) combined with effective SBCC will lead to improvements in children's and caregivers' diet. Since the SBCC activities were not yet widely rolled out, we cannot fully test this assumption. Instead, the impact evaluation can now only test whether the vouchers alone had a positive impact on diets.

62. However, the fact that only two-thirds of the households in the two voucher arms received vouchers and the fact that even fewer households received vouchers in the month before the endline evaluation means that testing the reduced evaluation question ("vouchers vs no vouchers") becomes difficult.

63. Our statistical power calculations (see above) assumed that all households in the voucher group will receive vouchers (and that none of the control households receive vouchers). Such non-compliance reduces the statistical power; i.e. our ability to detect whether the program had a positive impact on diets. This non-compliance may also reduce the inferences that can be drawn from the impact estimates, in particular if the households that received the vouchers have different characteristics than those households that were selected to the voucher arms but did not receive any vouchers.

64. Recall that we have set the minimum detectable effect sizes (MDEs) as follows:

1. Small voucher vs no-voucher: 15 percentage point difference in MAD prevalence
2. Large voucher vs no-voucher: 30 percentage point difference in MAD prevalence
3. Large voucher vs Small voucher: 15 percentage point difference in MAD prevalence

65. In our power calculations, we set significance level (α) at 0.05, power at 0.8 and accounted for intra-cluster correlation (0.03). The total number of clusters (villages) was set (after revision) to 60.

66. Low compliance has a strong negative effect on statistical power (Duflo et al. 2007).[6] As we note above in Table 5, 63 percent of the households in the small voucher arm received at least one voucher between June and December while the corresponding figure in the large voucher arm is 71 percent. We also saw that 16 percent of the non-voucher arm received a voucher. We can now calculate the compliance rate for each test involving the non-voucher group; Table 6.

Table 5: Percent of households receiving vouchers, by study arm

Test	Compliance rate	Statistical power
Small voucher vs no-voucher	0.63-0.16=0.47	0.38
Large voucher vs no-voucher	0.71-0.16=0.55	0.94

67. The last column of Table 5 reports the re-calculated statistical power once we take into account the compliance rates.[7] The statistical power in the first test (assuming the same

[6] For an accessible description of this issue, see https://blogs.worldbank.org/impactevaluations/power-calculations-101-dealing-with-incomplete-take-up.

[7] We also used the actual achieved sample sizes (see Section 4.2) in these calculations.

MDEs) is only 38 percent. This means that if the intervention was effective in reality, our test would be able to detect this only 38 percent of the time. In other words, our study is now severely under-powered to run the first test that compares the outcomes in the small voucher group against no-voucher group. The resulting statistical power is above 80 % for the second test implying that our study is still adequately powered to run the second test involving the larger voucher group and no-voucher group.

68. However, in the context that lacks technology to refrigerate the fresh foods[8], we should focus on the share of households receiving vouchers in the previous month preceding the interview. This is because most of our outcome variables capture dietary patterns in the 24 hours or 7 days preceding the interview. We therefore redo the power calculations considering the compliance rate in terms of receiving vouchers either in November or December. In the non-voucher group, 0.7 percent of the households received a voucher since November (see Table 4 above). The corresponding percentages for the small and large voucher groups are 10.4 percent and 11.1 percent, respectively. The resulting compliance rates and statistical power for each test are provided in Table 6.

Table 6: Percent of households receiving vouchers, by study arm

Test	Compliance rate	Statistical power
Small voucher vs no-voucher	0.097	0.065
Large voucher vs no-voucher	0.104	0.119

69. Now both statistical power estimates are well below 15 percent. With such low statistical power, our impact evaluation has a very low likelihood of detecting an impact of the program (should it exist). Perhaps a less appreciated aspect of low power is that it also reduces the likelihood that a statistically significant estimate reflects a true effect (Button et al. 2013).

70. In this setting, the impact evaluation would not yield reliable findings. Considering this, we refrain from reporting any impact estimates in this report.

Other limitations

71. Next we discuss the other limitations of this evaluation. First, due to the budget constraints, this evaluation focused on children 6-23m and their mothers (many of which are still lactating). Analyzing the impact on pregnant women would have required an additional sample. Apart from the considerable cost implications, there was also a

[8] Our baseline data indicate that only 1 percent of the households reported to own a refrigerator.

concern with the difficulty of finding sufficient number of pregnant women from the study area.

72. Second, we were not able to use quantitative research methods to answer Evaluation Question #4 ("What are the impacts of the project on the local markets of fresh foods?"). Answering this evaluation question using quantitative research methods would have required a sufficiently large sample of markets that are subject to the intervention and markets that are not (control markets). As the program is operating in 3-4 markets, there is not sufficient treatment sample available (let alone budget) to carry out this type of analysis. We therefore used qualitative research methods to study this evaluation question. This approach sheds some light on how the program changed the fresh food markets, but we cannot say anything definite about the program's impact on markets.

73. Third, and related, also the control households were likely to benefit from improved availability of fresh foods in the market. This would lead to positive spillover effects for the control households. On the other hand, if the supply of fresh foods in markets cannot keep pace with the increased demand (due to FFV), then we should expect to see the prices of these foods to rise. This would constitute a negative spillover effect. We used the qualitative surveys to assess the extent to which these issues are happening in the study area. However, it is not within the scope of this evaluation to assess these types of spillover effects in a more quantitative manner.

74. Fourth, the Fresh Food Voucher Pilot Programme was planned to be operating in 3 districts. Again, due to budget limitations, the evaluation focused only on one district (Habru). As a result, we tried to evaluate the impacts of the program in this district but cannot say whether the program works in the other districts.

75. Finally, the intervention – and consequently the study – area was purposely selected. The intervention focuses on markets that have the best capacity to absorb the increased demand induced by the vouchers scheme. This raises some concerns about external validity; the extent to which the results of this evaluation study can be generalized to the rest of Amhara (or Ethiopia). This is a particularly important question if the WFP (or the government of Ethiopia) plans to scale-up this FFV pilot programme.

Risks

76. We anticipated four risks with this evaluation. First, since we planned to re-interview the same households twice, there was some risk of respondent dropouts and non-response rates. This maybe because of the difficulty to trace the same respondent (respondents move and change address or die, or simply the proper address was not recorded). Or, it may be due to respondent fatigue such that the respondent is tired of getting involved in the same survey again. Our sample calculations accounted for 10 percent attrition.

We also devised a number of strategies and worked hard to minimize non-responses. In previous surveys, for example, in the five-round PSNP surveys between 2006 and 2014, IFPRI has been able to keep our non-response rates low by international standards.

77. Second, the village-cluster RCT design means that all eligible households within the same village receive the same treatment package (SBCC, SBCC + smaller voucher, or SBCC + larger voucher). As a result, we did not expect that the evaluation design causes discomfort among households that were not selected to receive vouchers. Still, to mitigate these risks, we recommended careful discussions and description of the evaluation approach with the communities before the implementation of the program began.

78. Third, contamination in randomized trials refers to a situation where the control households also directly benefit from the program. This would contaminate the control group invalidating the evaluation approach. In the context of FFV, we can think of two sources of contamination. The first source of contamination is that the voucher scheme is, by mistake, rolled out into the control villages. To minimize this from occurring, we carefully explained this issue to the implementation team to minimize this risk. The second potential source of contamination is that the beneficiary households will share a substantial part of their food with the control households. We thought that the chances of this happening are small given the clustered design: it is unlikely that the households in other villages know each other well, let alone learn each other's treatment status. Second, our previous work in the context of PSNP suggest that such sharing of transfers is rare in rural Ethiopia (Berhane et al. 2015). Still, mindful of this, the quantitative survey questionnaire included questions about the sources of households' food consumption. We can detect contamination by comparing the responses of the control households at the baseline to the responses at the endline. In the endline questionnaire, we added more questions on the use of vouchers to further assess the degree of contamination.

79. Fourth, the kind of political unrest that surfaced in 2016 in some parts of the country may delay survey – and implementation – activities. We had a proper contingency plan as to how we undertake the surveys without having this significantly influencing the project successes. Fortunately, the survey activities were not delayed because of these or any other reasons. However, as we discuss above, insecurity issues caused a significant delay to the *implementation* activities.

Ensuring quality

80. Dr Hirvonen (Team Leader) and Dr Baye (Deputy Team Leader) were be responsible for quality assurance. This evaluation had a number of in-built mechanisms to assure

quality. First, the survey instruments were first translated from English to Amharic and then back to English to be sure that all translations capture their intended meaning. Second, we carefully trained our enumerators and team members on how to conduct the field research. Moreover, before the actual fieldwork, we conducted a pre-testing that allowed us to detect any gaps in the enumerator training or problems with the survey instruments. Third, we closely collaborated with the district and sub-district officials to ensure that the field work minimizes the disruption caused to the local residents. Fourth, to minimize errors in the data entry phase, the data from the paper questionnaires were entered twice after which the results were compared, and inconsistencies solved. Fifth, to ensure transparency, we actively engaged with WFP and local stakeholders throughout the process. Finally, we are committed to making all research outputs from this project publicly available.

81. This evaluation satisfies the DEQAS criteria for Impartiality, Independence, Credibility, and Utility. The evaluation team is completely external and had no role in designing the intervention (**Independence**) and is free from any influence that may bias their reporting (**Impartiality**). The evaluation strategy is credible in the sense that is based on the most rigorous methods available and the evaluation team is committed to transparency throughout the evaluation process (**Credibility**). The evaluation will be of considerably use to decision-makers and stakeholders and the evaluation team is committed to making the results publicly available in a timely fashion (**Utility**).

Ethical issues

82. WFP's decentralized evaluations must conform to WFP and UNEG ethical standards and norms. The contractors undertaking the evaluations are responsible for safeguarding and ensuring ethics at all stages of the evaluation cycle. This includes, but is not limited to, ensuring informed consent, protecting privacy, confidentiality and anonymity of participants, ensuring cultural sensitivity, respecting the autonomy of participants, ensuring fair recruitment of participants (including women and socially excluded groups) and ensuring that the evaluation results in no harm to participants or their communities.

83. During the evaluation the following ethical issues were considered for the design, data collection, data analysis, reporting and dissemination.

84. First, this study does not have any significant physical, psychological, social, or economic risks. No medical device or therapy was part of this study. The methods of anthropometry and other survey measurements are non-invasive and virtually risk free. Respondents may find some items in the questionnaire modules to be of a personal nature, and they were be advised that they can choose not to respond to any questions

that make them feel uncomfortable, without ramifications for their participation in the study or their relationship with study staff or institutions.

85. Second, this evaluation complied with the code of U.S. federal regulations established by the Office of Human Research Protection (OHRP) and the international guidelines for ethical research follow standard practices regarding research involving human subjects.

86. Third, the consent process occurred in two phases. In the first phase, members of our survey supervision staff met with local leaders to describe the scope, purpose and duration of the study, the respondent burden, the potential risks and benefits and provide contact details of individuals in Ethiopia who can be contacted for additional details. In the second phase, enumerators described to both male and female household respondents the scope, purpose and duration of the study, the respondent burden, the potential risks and benefits and provide contact details of individuals in Ethiopia who can be contacted for additional details. It was stressed that participation was strictly voluntary and that participants can withdraw from the survey at any time. The survey team noted if voluntary consent was given. If consent was not given, enumerator did not proceed with the household interview.

87. Fourth, names and other easily recognizable identifiers were entered with household IDs in a separate sheet from all other data. This file containing names will be held separately from all other data files and will be kept only by the Principal Investigator (Hirvonen) and Co-Investigator (Baye). Finally, study identifiers (village and household IDs) are included in each data file so that data from the several instruments collected within a household may be linked together and with future survey rounds. However, these are not meaningful to casual observers without access to the original study logs. Public use data will include no identifiers.

88. Finally, prior to the beginning of the field work, the evaluation approach was reviewed and approved separately by the Institutional Review Boards (IRB) of IFPRI, Addis Ababa University and the Amhara regional review board.

2. Evaluation Findings

89. The evaluation findings and the evidence to substantiate them are presented below. They are structured as a response to each evaluation question in turn.

2.1. What are the differential impacts of the programme on diet diversity for the different voucher values?

90. The lack of implementation fidelity (see Section 1.3) does not allow the evaluation team to answer this question.

91. Below we report the changes in the outcome variables related to food security and nutrition.

Food security [9]

92. WFP has developed a Food Consumption Score that is specifically for the analysis of nutritional quality (WFP 2015). The FCS-N indicator focuses on the consumption of foods that are rich in protein, vitamin A and hem-iron.

93. Table 7 shows the frequency of consumption of these foods. We see that the consumption of protein rich foods is common in Habru. In both base and endline, the average household consumed protein rich foods nearly every day of the week (6.8 days). In contrast, the consumption of vitamin A and hem iron rich foods is less frequent. At the baseline, the average household consumed vitamin A rich foods on 1.4 days and at the endline on 2.1 days. The corresponding numbers for hem iron rich foods is 0.3 days both at the baseline and baseline.

94. We also collected information on additional food security indicators, such as WFP's Food Consumption Score (FCS) and Household Food Insecurity Access Scale (HFIAS) that were not the main outcomes of interest. The findings based on these indicators are reported in Annex 6.

Table 7: Number of days households consumed Protein, Vitamin A and Hem iron rich foods, by survey round and study arm

	baseline	endline	difference
Protein rich foods (number of days)			
Non-voucher	6.9	6.9	0.0
Small voucher	6.8	6.8	-0.1
Large voucher	6.6	6.7	0.0
Total	6.8	6.8	0.0
Vitamin A rich foods (number of days)			
Non-voucher	1.6	2.1	0.6
Small voucher	1.6	2.2	0.6
Large voucher	1.2	2.0	0.8
Total	1.4	2.1	0.7
Hem iron rich foods (number of days)			
Non-voucher	0.3	0.3	0.0
Small voucher	0.3	0.4	0.0
Large voucher	0.3	0.3	-0.1
Total	0.3	0.3	0.0

[9] Note that some of the baseline statistics reported here differ from those reported in the baseline report. This is because we have re-calculated them using the survey weights described in section 1.3.

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3 ; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

Diets [10]

95. The quantitative survey instrument contained a series of questions on household, child and caregiver diets. Following Swindale and Bilinsky (2006), the household level dietary module was based on a 7-day recall, and assessed using 12 food groups.[11]

96. Table 8 lists these food groups and shows the percent of households consuming from each of them. As expected, virtually all households consumed cereals and pulses. At the baseline, 28 percent of the households reported to have consumed roots/tubers in the past 7 days and this has increased to 55 percent at the endline. As for fresh foods, at the baseline, more than 76 percent of the households consumed vegetables while only 17 percent reported consuming fruits. We see some improvements in the endline: 82 percent of households consumed vegetables and 44 reported consuming fruits. Meanwhile, we see little improvement in the consumption of animal source foods; only 3 percent of the households in both rounds reported to have consumed meat/poultry, about 20 percent consumed eggs and about 25 percent of households consumed dairy in the 7 days prior to the interview.

Table 8: Percent of households consuming from each HDDS food group

	Baseline				Endline			
	N0-V	**SV**	**LV**	**All**	**N0-V**	**SV**	**LV**	**All**
Cereals	100	100	100	100	100	100	99	100
Roots/tubers	36	27	21	28	53	56	56	55
Vegetables	87	78	65	76	85	86	73	82
Fruits	21	20	10	17	47	46	39	44
Meat/poultry	3.6	2.8	3.2	3.2	5.2	3.0	2.0	3.4
Eggs	24	22	20	22	25	20	17	21
Fish *	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nuts/pulses	100	100	99	100	99	97	96	97
Dairy	29	22	22	25	33	28	24	28
Oils/fats	92	92	81	88	91	92	88	91
Sugar/honey	35	28	23	29	67	55	47	56
Coffee/tea	82	82	75	80	91	91	88	90

*Note: * Fish consumption frequency is assumed zero. No-V refers to non-voucher arm, SV to small voucher arm and LV to large voucher arm.*

[10] Note that some of the baseline statistics reported here differ from those reported in the baseline report. This is because we have re-calculated them using the survey weights described in section 1.3.

[11] The only difference is that we did not ask about fish consumption. The decision to not ask about fish consumption was based on our previous work in the area, according to which practically no one consumes fish.

97. Table 9 shows the mean HDDS by study arm and survey round. At the baseline, the average household in this sample consumed from 5.7 food groups (out of the maximum 12) while in the endline, this has increased to 6.8 food groups. We see that household dietary diversity increased in all study arm approximately by one food group; differences between control and voucher arms are not statistically different from zero.

Table 9: Mean HDDS, by study arm and round

Study arm	Baseline	Endline	Difference
Non-voucher	6.0	6.9	0.9
Small voucher	5.7	6.8	1.1
Large voucher	5.5	6.6	1.1
All households	**5.7**	**6.8**	**1.1**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3 ; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

98. Diets of the caregivers were assessed using a 24-hour recall. Following FAO and FHI 360 (2016), the responses to these questions were then grouped into 10 food groups (listed in Table 10 below).

99. Table 10 shows that the dietary patterns are similar to what was observed at the household level; cereals and pulses dominate while the consumption of (other) vegetables is also quite common. The consumption of other food groups is much less common. Compared to the baseline, we see small increase in the share of women that consumed grains, roots and tubers (4 percentage points) and (other) vegetables (11 percentage points).

Table 10: Percent of female caregivers consuming from each MDD-W food group

	Baseline				Endline			
	No-V	**SV**	**LV**	**All**	**No-V**	**SV**	**LV**	**All**
Grains, roots, and tubers	96.4	93.6	94.0	94.7	97.0	99.5	98.5	98.3
Pulses	92.9	86.2	92.6	90.5	83.0	90.6	87.8	87.1
Nuts and seeds	9.3	6.9	6.0	7.4	0.7	3.0	3.6	2.4
Dairy	6.4	6.5	7.9	6.9	14.1	6.4	10.2	10.2
Meat, poultry, and fish	3.6	2.3	1.9	2.6	4.4	2.5	2.5	3.2
Eggs	0.0	0.5	0.0	0.2	0.7	1.5	0.5	0.9
Dark leafy green vegetables	2.1	0.9	0.5	1.2	5.9	3.0	4.1	4.3
Other Vitamin-A rich fruits	3.6	0.0	1.4	1.7	0.7	0.5	3.0	1.4
Other vegetables	77.1	72.4	60.6	70.0	87.4	83.7	72.1	81.1
Other fruits	0.0	0.9	0.9	0.6	5.2	5.4	4.6	5.1

Note: No-V refers to non-voucher arm, SV to small voucher arm and LV to large voucher arm.

100. At the baseline, the mean dietary diversity in this sample was low at 2.8 food groups (Table 11). Only 0.4 percent of the women met the minimum recommended dietary diversity (5 out of 10 food groups). The women in the two voucher arms had lower dietary diversity score than women in the control arm (p<0.05). At the endline, we see a marginal improvement in women's dietary diversity score (2.9 food groups). This time, 3 percent of the women met the minimum recommend dietary diversity.

101. None of the differences between the control arm and the voucher arms are statistically significant at the endline. The changes in the dietary diversity score were larger in the small voucher arm compared to the control arm (p<0.05). However, the differences in the change in the main outcome variable – the share of women meeting MDD-W – were not statistically different from zero.

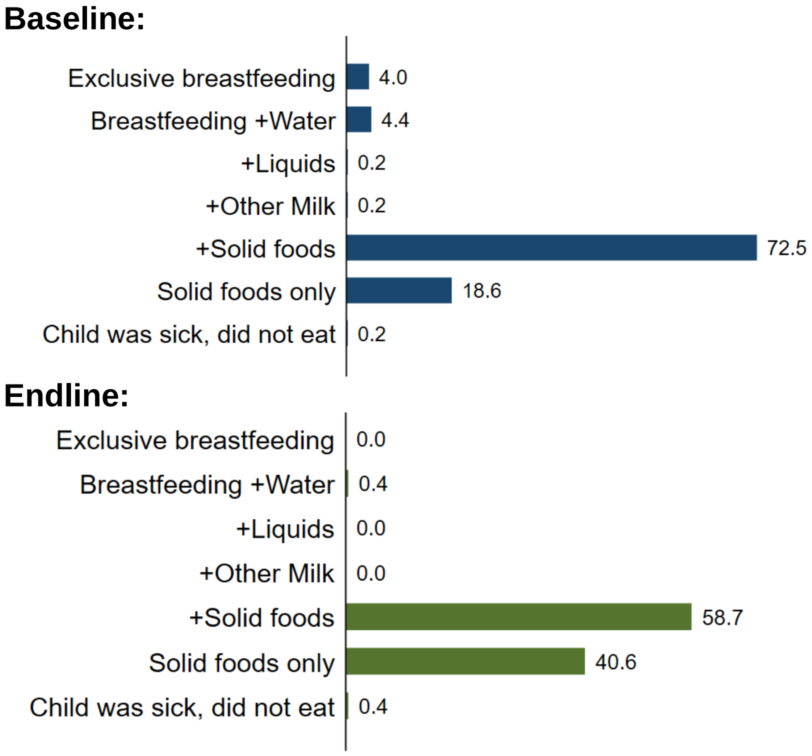
Table 11: Mean female caregiver dietary diversity score and MDD-W, by study arm and survey round

Study arm	Baseline	Endline	Difference
Number of food groups			
Non-voucher	2.9	2.9	0.0
Small voucher	2.7 ***	3.0	0.3 **
Large voucher	2.7 **	2.9	0.2
All mothers	**2.8**	**2.9**	**0.2**
Minimum Diet Diversity for Women (%)			
Non-voucher	0.7	2.7	2.0
Small voucher	0.8	2.5	1.8
Large voucher	0.0	3.6	3.6
All mothers	**0.4**	**3.0**	**2.6**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3 ; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

102. The attention now shifts to children's diets. We begin by noting that at the baseline, children were 6-17 months of age and consequently, at the endline (administered 12 months later), the age range is 18-29 months. This aging of the sample has implications to the breastfeeding and complementary feeding patterns. As shown in Figure 7, while 4 percent of the children were exclusively breastfed at the baseline, all children are introduced to complementary foods by the endline. Moreover, 81 percent of the children at the baseline were breastfed at the time of the survey. At the endline, this percent has decreased to 59 percent. The other implication of the aging of the sample is that some children are older than 23 months of age. Most of the child feeding indicators have been developed, and validated, for 6-23 month old children. In what follows, we have decided to keep the older children in the sample to maximize the sample size.

Figure 7: Child feeding status (%), by survey round



103. The survey instrument also asked about children's food consumption in the last 24 hours. Following WHO (2008) guidelines, children's food consumption was categorized into 7 food groups. Table 12 lists these food groups and shows the percent of children consuming from each of them. As before, food consumption is concentrated on cereals and legumes. Compared to the baseline, the share of children consuming from the food group has increased across almost all food groups. This is expected as all children are now introduced to complementary foods and also the variety of foods children consume increases by age. Compared to the baseline, the share of children consuming fruits and vegetables has increased. At the endline, nearly 10 percent of the children consumed vitamin A rich fruits and vegetables, up from 3 percent at the baseline. Similarly, 73 percent of the children consumed fruits and vegetables that are not rich in Vitamin A at the endline, up from 43 percent at the baseline.[12] With the exception of dairy, the consumption of animal source foods is rare with little changes between the two survey rounds.

Table 12: Percent of children consuming from each food group, by survey round and study arm

	Baseline	Endline

[12] It is worth reminding here that these changes are not driven by seasonality: the 2017-baseline and the 2018-endline took place at same time of the year (December).

	No-V	SV	LV	All	No-V	SV	LV	All
Grains, roots and tubers	87.1	87.8	81.8	85.6	97.0	98.5	97.5	97.7
Legumes and nuts	56.4	62.0	63.1	60.5	76.9	84.7	82.2	81.3
Dairy products	27.1	24.4	22.4	24.7	26.1	22.8	18.8	22.6
Flesh foods (red meat, poultry, seafood)	1.4	1.9	2.8	2.0	3.7	4.0	2.0	3.2
Eggs	5.7	3.3	2.3	3.8	2.2	5.0	2.5	3.2
Vitamin A rich fruits and vegetables	2.1	2.8	3.7	2.9	11.2	8.4	9.1	9.6
Other fruits and vegetables	42.1	49.8	35.5	42.5	79.1	74.3	66.0	73.1

Note: No-V refers to non-voucher arm, SV to small voucher arm and LV to large voucher arm.

104. As shown in Table 13, at the baseline, the average child in this sample consumed from 2.4 food groups and 13 percent met the recommended dietary diversity (4 food groups out of 7). At the endline, the mean dietary diversity score has increased to 3.1 and now 22 percent of the children meet the recommended dietary diversity.

105. At the baseline, the children in the non-voucher arm were more likely to have met the recommended dietary diversity than the children in small voucher arm ($p<0.10$). By the endline, these differences between the study arms have narrowed, and are no longer statistically different from zero. However, the increase in the share of children meeting minimum dietary diversity was larger in small voucher arm as compared to control arm ($p<0.10$).

Table 13: Mean child dietary diversity score and % of children meeting the recommended dietary diversity, by study arm and survey round

	Mean dietary diversity score			% of children meeting the min dietary diversity		
Study arm	**Baseline**	**Endline**	**Difference**	**Baseline**	**Endline**	**Difference**
Non-voucher	2.5	3.2	0.7	16.4	20.7	4.2
Small voucher	2.3	3.2	0.8	10.1 *	23.2	13.1 *
Large voucher	2.4	3.1	0.7	12.2	22.6	10.4
All children	**2.4**	**3.1**	**0.8**	**12.6**	**22.3**	**9.7**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

106. The definition of the minimum meal frequency depends on the age of the child and his/her breastfeeding status (WHO 2010):

- Breastfed infants 6-8 months of age: at least 2 feedings of solid, semi-solid, or soft food in the last 24 hours
- Breastfed infants 9-23 months of age: at least 3 feedings of solid, semi-solid or soft foods in the last 24 hours.
- Non-breastfed children, 6-23 months of age: at least 4 or more feedings of solid, semi-sold or soft foods in the last 24 hours.

107. Table 14 shows the share of children meeting the recommended meal frequency for study arm and by survey round. At the baseline, 69 percent of the children met the minimum meal frequency as recommended by the WHO. At the endline, 87 percent of the children meet the minimum meal frequency.

108. While there were not statistically significant differences in baseline percentages, at the endline, children in the two voucher arm are more likely to have met the minimum meal frequency (p<0.05). The last column of the table tells us that the increase in the share of children meeting this target is larger in the two voucher arms as compared to the control arm. However, the difference to the control arm is only statistically significant in the small voucher arm (p<0.01).

Table 14: % of children meeting the recommended meal frequency by study arm and round

Study arm	Baseline	Endline	Difference
Non-voucher	70.7	79.6	8.8
Small voucher	64.3	88.1 **	23.8 ***
Large voucher	72.2	90.6 ***	18.4
All children	**69.2**	**86.8**	**17.6**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

109. Finally, the minimum acceptable diet (MAD) is a composite indicator of the minimum dietary diversity (MDD) and minimum meal frequency (MMF) and calculated for children between 6 and 23 months of age. As with MMF, the definition of MAD depends on the breastfeeding status of the child:

- Breastfed children: achieve the minimum dietary diversity AND the minimum meal frequency;
- Non-Breastfed children: achieve the minimum dietary diversity AND the minimum meal frequency AND received at least 2 milk feedings.

110. At the baseline, only 8 percent of the children met MAD (Table 15) while at the endline this percentage has increased to 22 percent. At the baseline, the differences between the control arm and the two voucher arms are not statistically different from zero. The same is true for the endline. Finally, the last column shows the change in MAD prevalence between the baseline and endline. While the improvements are larger in the two voucher arms as compared to the control arms, these 'differences in differences' are not statistically significant.

Table 15: % of children meeting the minimum acceptable diet (MAD) by study arm and round

Study arm	Baseline	Endline	Difference
Non-voucher	10.8	17.9	7.2
Small voucher	6.0	21.8	15.8
Large voucher	8.6	25.5	17.0
All children	**8.3**	**22.3**	**14.0**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

111. Table 16 disaggregates the key child dietary outcomes by child's sex. At the baseline, nearly 8 percent of the boys and 4 percent of the girls in our sample met the criteria for minimum acceptable diet. This difference is statistically different from zero at the 10-

percent level. At the endline, the order has reversed. In December 2018, 22 percent of the girls and 18 percent of the boys met the minimum acceptable diet. However, this difference is not statistically different from zero, indicating the gap observed in the baseline has closed. Finally, the endline data also tells us that compared to boys, girls were less likely to have met the minimum dietary diversity but more likely to have met the minimum meal frequency.

Table 16: % of children meeting MAD, min dietary diversity and meal frequency by study arm and round

	Baseline		Endline	
Diet indicator	**Female**	**Male**	**Female**	**Male**
% meeting the min acceptable diet	4.2	7.7	22	18
% meeting the min dietary diversity	9.1	12	20	23
% meting the min meal frequency	66	70	88	86

Key findings and conclusions

The lack of implementation fidelity does not allow the evaluation team to answer this question.

2.2. What are the main changes in knowledge, attitude and practices of the beneficiary households regarding access and use of nutritious foods?

112. The SBCC activities were not yet rolled out at the time of the endline survey. Consequently, there is little change in knowledge and attitudes.

113. All caregivers of the index children were asked about their knowledge and attitudes regarding infant and young child feeding (IYCF) practices. Table 17 shows the questions and the percent of caregivers who provided a correct answer to the question. We see that the knowledge on the age in which different types of foods should be initiated is poor. A closer look at the responses reveal that caregivers think that the introduction of various nutritious foods should be delayed. Moreover, misconceptions exist with respect to optimal frequency in which various nutritious foods are given to the child.

Table 17: % of correct responses to the nutrition knowledge questions

Correct response (%)	baseline	endline
Age when baby should start receiving liquids other than breast milk?	81	87
Age when baby should start eating animal source foods?	53	61
At what age should a baby first start to eat fruits?	40	40
At what age should a baby first start to eat vegetables?	39	40
Common problem with traditionally prepared gruels given as first foods?	18	12
How often should a baby eat animal foods such as eggs and milk?	90	87
How often should a baby eat fruits?	43	78
How often should a baby eat vegetables?	38	57
What are some of the foods that contain vitamin A?	39	74
How often should a child be fed when he/she is sick?	52	56

114. We then aggregated these responses into a nutrition knowledge score. Each correct response was given one point, yielding an IYFC score ranging between 0 and 10. The mean score among the caregivers at the baseline was 5.1 (median: 5) and at the endline 5.9 (median 6). Table 18 disaggregates the mean scores further by study arm and survey round. [13] The knowledge levels in each survey rounds are similar and not statistically different from zero across all study arms as are the changes between the two rounds.

Table 18: The mean caregiver nutrition knowledge score by study arm and round

Study arm	Baseline	Endline	Difference
Non-voucher	5.19	6.21	1.01
Small voucher	5.03	5.80	0.77
Large voucher	5.06	5.64	0.58
All households	**5.09**	**5.85**	**0.76**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

115. The qualitative data collected at the baseline support the above findings. As such, no taboos were identified against the consumption of fresh foods, but there were practical concerns (e.g. choking). Many caregivers believe that certain foods (e.g. kale) should be introduced late, after the child's first birthday. The reasons for the delay relate to misconceptions regarding difficulty for the child to digest the food, and food safety particularly for uncooked fresh foods. This contrasts with caregivers' perceived knowledge as they consider themselves knowledgeable, thanks to the increased awareness raised by the routine nutrition education they receive. This theme persistent

[13] Note that some of the baseline statistics reported here differ from those reported in the baseline report. This is because we have re-calculated them using the survey weights described in section 1.9.

in the endline FGDs where many respondents identified poverty, not the lack of knowledge as the main constraint. As mother of one girl in Kule Koba responded:

"I try to buy one or two oranges for my children -at least in one of the two weekly markets- but, I cannot buy fruits all the time…"

116. Others supported the idea and one mother of two children who had recently returned from Saudi Arabia:

"It is not that we do not know the importance of fruits and vegetables to the health of our children…. I personally have the exposure, as I have witnessed what children consume in Saudi Arabia. However, I am not able to do the same here, because it is not affordable, and also because we do not have refrigeration."

117. This perception and the actual gap in knowledge should be taken into consideration in the SBCC implementation.

118. As for practices, we can study the purchasing patterns among households that did receive FFV transfers. The quantitative survey instrument asked voucher-households how they used their voucher. Table 19 shows the share of households purchasing different food items with the last voucher they received. We see that all voucher households purchased at least one vegetable with almost all households purchasing Irish potatoes, sweet potatoes and onions. Households that received the larger valued voucher were somewhat more likely to purchase sweet potatoes, tomatoes, spinach and carrot than households that received a smaller voucher.

119. About 95 percent of the households purchased fruits, typically bananas, with their voucher. Compared to households receiving the lower valued vouchers, households in the large voucher arm were considerably more likely to purchase mangoes and oranges with their transfers. As for animal source foods (ASFs), we see that only 31 percent of the households purchased these with their voucher, possibly reflecting the limited supply of these products among the FFV traders. The most common purchased ASF was eggs. Moreover, the differences in ASF purchasing patterns between the two voucher groups are small.

Table 19: Percent of voucher households purchasing different food items

	Both groups	Small voucher	Large voucher	Difference
Any vegetable	**100**	**100**	**100**	**0**
Irish potato	98	98	99	2
Sweet potato	76	71	80	9
Beetroot	47	48	46	-2
Onion	98	98	99	0

Tomato	59	54	63	9
Kale	52	51	52	1
Lettuce	12	11	12	1
Spinach	3	1	6	5
Carrot	33	31	34	3
Other vegetable	44	47	41	-6
Any fruit	**95**	**94**	**96**	**2**
Mango	13	7	17	10
Papaya	2	0	3	3
Banana	83	83	82	-1
Orange	55	42	65	23
Avocado	3	1	4	4
Other fruit	7	8	6	-2
Any ASF	**31**	**30**	**31**	**1**
Egg	21	22	21	-1
Beef	5	4	6	2
Goat meat	9	8	10	2
Chicken meat	0	1	0	-1

Note: Sample restricted to 269 households that belong to the two voucher arms and received transfers; 128 households in the small voucher arm and 141 households in the large voucher arm.

120. The qualitative interviews with the caregivers confirm these purchasing patterns. Caregivers that participated in the FGDs reported to prefer buying products with better shelf-life like onions and potatoes. This was largely motivated by limited access to preservation techniques combined with the fact that markets take place only once or twice a week.

Key findings and conclusions

The SBCC activities were not yet rolled out at the time of the endline survey. Consequently, there is little change in knowledge and attitudes. However, in terms of practices, the descriptive and qualitative data provide some suggestive evidence that among beneficiary households, FFV has improved access of fresh foods, particularly fruits and vegetables. These changes were however mostly observed by an increase in the consumption of onion and potatoes – items belonging to food groups which were widely consumed already prior to the intervention. The primary reason for selecting these food items relates to their better storability.

2.3. Which transfer value is more cost-effective in delivering nutritional results?

121. To answer this question we need estimates of the impact of the program on nutritional outcomes. However, the lack of implementation fidelity means that (see Section 1.3), the impact evaluation could not be carried out.

Key findings and conclusions

The lack of implementation fidelity does not allow the evaluation team to answer this question.

2.4. What are the impacts of the project on the local markets of fresh foods?

122. As described in Chapter 2, this research question was assessed mainly using qualitative data. In the FGDs with the traders, all participants agreed that they have witnessed an increased demand for fruits and vegetables, but less so for ASFs. The increased demand was reasonable and did not create any difficulties in ensuring sufficient supply. The main challenge was predicting the demand, as the flow of FFV transfers was not predictable. Indeed, the discussants unanimously said that the voucher system was not working well in the last months. An FFV trader (male, in his 30s) noted:

> *"We used to consider ourselves as special traders, but we have recently become a regular trader as we do not have many voucher recipients coming, because they did not get their transfers…"*

123. ASF traders complained about the low demand for ASFs. Eggs are bought by transfer recipients at about the time they receive their transfers and then the demand drops immediately. Because there is no organized (pre-) order for meat, the traders reported that they have stopped slaughtering as they are not sure of the demand.

124. Another challenge was raised from FGDs with traders in Dire Roka market. The beneficiaries in Dire Roka were perceived as not knowing how to cook vegetables. Traders mentioned that cooking demonstrations and education for beneficiaries would be good. The following were additional points identified as challenges by the FFV traders:

- Transfer delay and inconsistency;
- Coordinator not having enough information about the program (e.g. transfer timing);
- Beneficiaries do not know how to use the Hello-cash; traders said that they have to spend additional time to help them.

125. The market observations revealed that the FFV stalls had a hanging poster that allowed FFV beneficiaries to easily identify their traders. The FGDs with the traders revealed that while vegetables and fruits were widely available, traders of animal source foods like eggs and meat were rarely seen.

126. All FFV traders that took part in the FGDs, agreed that they have benefited, but could have benefited more if recipients received consistent support from the FFV program. They did not think other traders were hurt. Non-FFV traders argued that they sell better quality products and at a lower price than FFV traders.

127. The FGDs with the caregivers indicated that some FFV beneficiaries were not happy with the quality and prices of products sold by FFV traders. Mother in Dire Roka commented:

"I received the transfer only ones, but even then I could not get good quality products in the market".

128. Another woman followed:

"You can buy 5 bananas with 10 birr in non-FFV traders, while the same will cost 30 birr in FFV traders"

129. The caregivers that benefited from the FFV program requested WFP to carefully monitor the prices and the quality of produces being sold by the FFV traders.

Key findings and conclusions

- The qualitative evidence gathered from the FGDs with the traders suggests that project improved the availability of more fruits and vegetables in the markets. However, some caregivers (both beneficiaries and non-beneficiaries) and non-FFV traders complained that the FFV traders sell the same fresh foods at a higher price.

2.5. Gender issues

130. Although not explicitly mentioned in the evaluation matrix of the ToR, the evaluation team assessed various gender issues with the data that was collected. First, at the baseline, we collected data on fathers' (or male caregivers) dietary diversity and compared the maternal diets to those of the fathers. To do so, we restricted the analysis to 498 households in which both mother and father were present and ate at home in the last 24 hours. Table 20 compares the dietary outcomes between men and women in these households. The mean dietary diversity score among women was 2.73 and among men 2.76. Men were more likely to meet the minimum recommended dietary diversity. However, neither of these differences are statistically different from zero. The table also reports the proportion of men and women consuming from different food groups. These data reveal that the dietary content is very similar between men and women.

Table 20: Comparison of dietary outcomes between men and women

Food group	Women	Men
Dietary diversity score (based MDD-W)	2.73	2.76
% meeting minimum recommended dietary diversity	0.6	1.2
Grains, roots, and tubers	94.2	95.6
Pulses	90.9	91.0
Nuts and seeds	7.4	7.6
Dairy	7.6	8.6
Meat, poultry, and fish	2.6	1.6
Eggs	0.0	0.4
Dark leafy green vegetables	1.0	1.4
Other Vitamin-A rich fruits and vegetables	1.4	1.0
Other vegetables	67.2	67.5
Other fruits	0.8	0.8

Note: sample restricted to households where both female and male caregiver are present and ate at home in the last 24 hours; N=498; baseline data only.

131. Second, our quantitative survey instrument asked the mothers about the role of men in child feeding. In about two-thirds of the households, mother and father discussed about how to feed children. In 62 percent of the households, father is directly involved or supported the mother to feed the child. Typically, this meant that the father fed the child himself (29 % of all households), provided money to purchase nutritious foods, such as animal sourced foods, to young children (22 %), or helped by doing other household tasks while the mother was feeding the child.

132. Finally, the qualitative FGDs administered in the baseline and endline tried to capture the role of men in market participation. Our results reveal that there is little engagement by men in the market. Asking about decision-making regarding consumption within the households, we observed that in most cases, women are the ones making those decisions. This was illustrated in a response of a mother of two children in Kule Koba:

> *"What else should I know if I do not know what my husband and children want to eat…"*

3. Conclusions and Recommendations

133. Based on the findings presented in the previous section, an overall assessment that responds to the evaluation questions is provided below. This is followed by recommendations of how WFP can take action to build on the lessons learned.

3.1. Overall Assessment/Conclusions

134. The evaluation team was requested to answer four questions listed in the TOR:

Q1. What are the differential impacts of the programme on diet diversity for the different voucher values?

Q2. What are the main changes in knowledge, attitude and practices of the beneficiary households regarding access and use of nutritious foods?

Q3. Which transfer value is more cost-effective in delivering nutritional results?

Q4. What are the impacts of the project on the local markets of fresh foods?

135. Unfortunately, due to lack of implementation fidelity many of these questions could not be answered. In the impact evaluation literature, implementation fidelity is defined as the degree to which the implementation of the program adheres to the original design. Lack of implementation fidelity is typically the main reason why even theoretically sound programs fail to show positive impacts. The theory of change of the FFV program relies

on two interventions: SBCC activities and the delivery of fresh food vouchers. This endline evaluation began by examining the implementation of these two core activities.

136. The SBCC activities were severely delayed. The quantitative and qualitative data show that these activities were not yet rolled out at the time of the endline in December 2018.

137. Regarding the dissemination of the fresh food vouchers, the payment cycle data provided by the WFP shows that the first vouchers were only dispatched on 26 June 2018 indicating a 6-month delay. After this, the payment cycle followed a bi-monthly – instead of the planned monthly – schedule. Moreover, the quantitative data showed that about one-third of the households selected into the two voucher arms did not receive any vouchers between June and December. Meanwhile, 16 percent of the households in the non-voucher (control) arm received a voucher. If we consider the month preceding the endline survey, we see that only about 11 percent of the households in the voucher arms received a voucher while less than 1 percent of the households in the non-voucher arm received a voucher. The caregivers selected for the FGDs highlighted similar implementation challenges with respect to the payments. While the FFV was positively perceived by many caregivers, the inconsistency of the transfers and the limited information and support available was a source of complaint. The evaluation team was unaware of these transfer related implementation issues before the endline surveys began and therefore could not react to this, for example, by postponing the endline survey.

138. These implementation challenges mean that the impact evaluation could not be carried out. Our statistical power calculations assumed that all households in the voucher group will receive vouchers (and that none of the control households receive vouchers). This low compliance reduces statistical power; i.e. our ability to detect whether the program had a positive impact on diets. Consequently, the impact evaluation would not yield reliable findings and therefore, we refrained from reporting the impact estimates.

139. Finally, the terms of reference (ToR) for this evaluation had four questions that we list below, together with our responses to them in the light of the findings reported in this endline report:

140. *Q1. What are the differential impacts of the programme on diet diversity for the different voucher values?*

- Unfortunately, the lack of implementation fidelity does not allow the evaluation team to answer this question. Descriptive analysis of the voucher purchasing patterns suggest that the larger voucher somewhat increased the likelihood that the households purchased more expensive fruits, such as mangoes and oranges that are rich in Vitamin A and C. However, no notable differences were found with respect to animal source foods, possibly because of the limited supply of these products among the FFV traders.

141. *Q2. What are the main changes in knowledge, attitude and practices of the beneficiary households regarding access and use of nutritious foods?*

- The SBCC activities were not yet rolled out at the time of the endline survey. Consequently, there is little change in knowledge and attitude, but in terms of practice, the descriptive data provide suggestive evidence that FFV improved access to fresh foods, particularly fruits and vegetables. These changes were however mostly observed by an increase in the consumption of onion and potatoes – items belonging to food groups which were widely consumed already prior to the intervention. The primary reason for selecting these food items relates to their better storability.

142. *Q3. Which transfer value is more cost-effective in delivering nutritional results?*

- Unfortunately, the limited implementation fidelity does not allow the evaluation team to answer this question.

143. *Q4. What are the impacts of the project on the local markets of fresh foods?*

- The evidence gathered from the qualitative interviews suggest that the project has improved the availability fruits and vegetables in the markets. However, there are fears that the prices have also increased after the FFV was introduced in these markets.

144. Although not explicitly mentioned in the evaluation matrix of the ToR, the evaluation team assess various gender issues with the data that were collected. The findings from these analyses can be summarized as follows. First, the differences in children's dietary outcomes did not vary markedly by child's sex. Second, the dietary diversity between male and female caregivers were nearly identical before the intervention began. Third, in more than 60 percent of the household, the male caregiver is directly involved and supported the mother to feed child. Finally, analysis of the qualitative data reveal that there is little engagement by men in the market. Asking about decision-making regarding consumption within the households, we observed that in most cases, women are the one making those decisions.

3.2. Recommendations

145. Based on the findings and conclusions of this evaluation, the recommendations of the evaluation team are outlined below. The target group for each recommendation is clearly identified. The recommendations are structured by priority.

Recommendation 1: The WFP needs to make sure that all eligible households receive transfers on a monthly basis and that the transferred amounts are both predictable and consistent across months.

Recommendation 2: The WFP needs to hire dedicated program coordinators, if possible, at the kebele level to provide on-the-spot support to beneficiaries and eligible households that are not receiving vouchers. A complaint feedback mechanism that complementing the hotline number should be set-up.

Recommendation 3: The WFP should reconsider excluding these onions and potatoes from the voucher scheme to maximize the dietary quality impacts.

Recommendation 4: The WFP needs to make sure that all beneficiary households receive SBCC. The gap between perceived nutrition knowledge and actual nutrition knowledge need to be taken into account in the SBCC strategy.

Recommendation 5: The WFP should consider providing specific training on how to cook and preserve fresh foods.

Recommendation 6: The WFP needs to closely monitor the intervention markets to ensure that the quality of the foods supplied by the FFV vouchers and their prices remain reasonable.

Annexes

Annex 1: SBCC strategy

The FFV program includes a large behavioural component where beneficiaries are given information about the importance of a diverse diet, along with other social behaviour change communication (SBCC) initiatives such as practical demonstrations on the proper usage of nutritious foods. The SBCC is led by an independent consultant, Dr Rowena Merritt, who has extensive expertise in designing SBCC strategies. The objectives of the SBCC strategy are 7-fold:

i. To create awareness of nutritional choices and the positive impact good nutrition can have on personal and child health and development.

ii. To improve knowledge about how to make positive changes to nutritional health.

iii. To increase the number of mothers and mothers-to-be who believe that a diet of fresh produce is ‘for people like them’.

iv. To provide mothers and mothers-to-be with practical cooking skills (i.e. how to add fresh produce into their staple diet).

v. To increase the amount of fruit and vegetables consumed by pregnant and lactating women, and children aged 6-23 months.

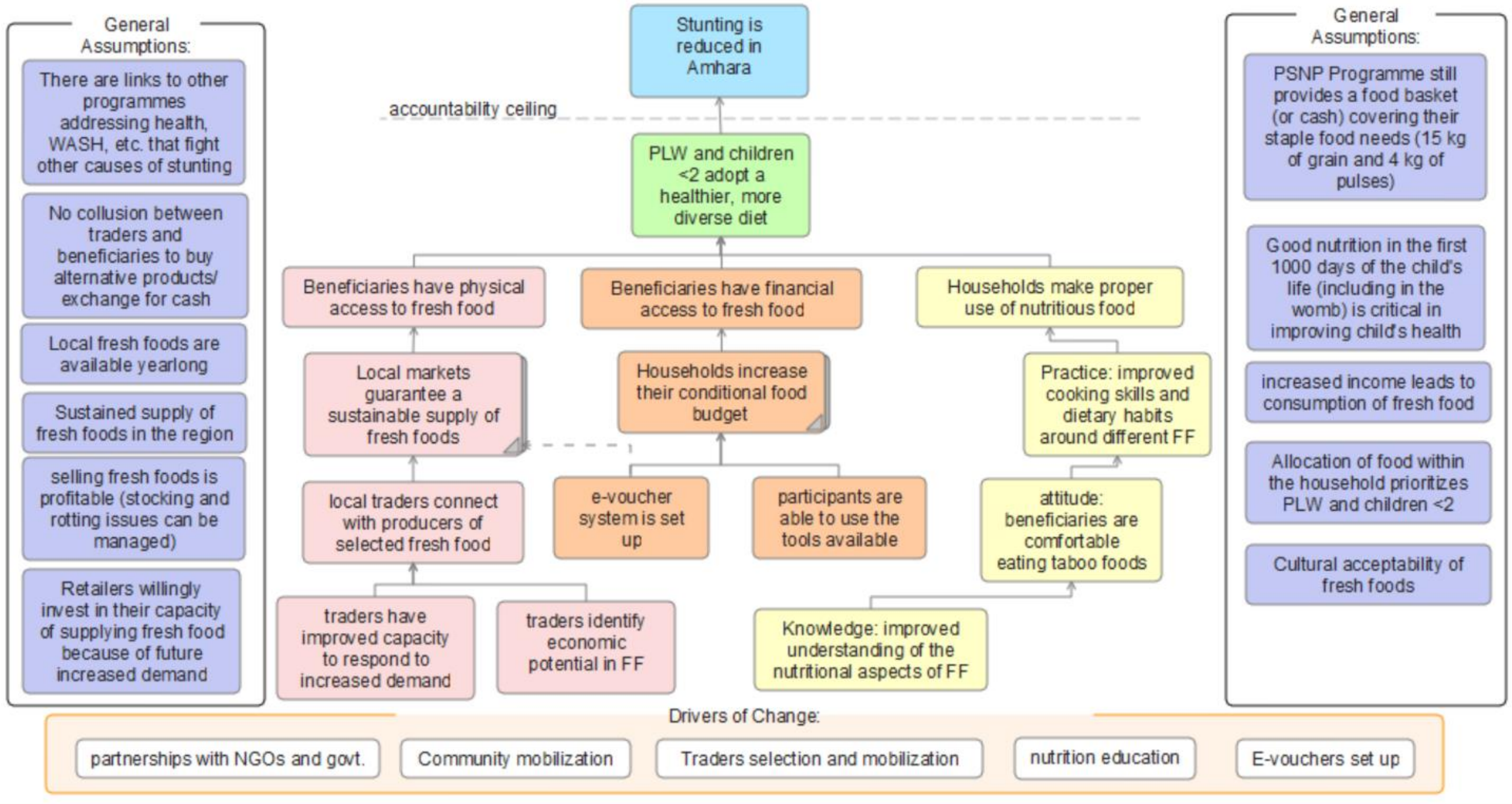
vi. To increase the amount of protein products consumed by pregnant and lactating women, and children aged 6-23 months.

vii. To increase the variety of fresh fruit and vegetables consumed by pregnant and lactating women, and children aged 6-23 months.

The SBCC strategy includes four main activities. First, the core products to provide practical SBCC material and resources were developed and/or revised. The target audience was directly involved in the development of these resources. Once finalized, the material (e.g. leaflets, posters) was disseminated in the communities. Second, in terms of SBCC activities, the SBCC component included a comedy community theatre, ‘coffee conversation’ events and cooking demonstrations to directly promote appropriate Infant and Young Child Feeding (IYCF) practices. Third, the Health Extension Workers (HEWs) are trained in Motivational Interviewing (i.e. create an environment of peer-to-peer support as opposed to a teacher-student relationship and build self-efficacy amongst the women). Fourth, the SBCC engages with religious leaders to promote the message around exemptions during fasting.

These SBCC activities are rolled out in all project localities and are not conditional on households receiving the voucher. Furthermore, the SBCC messages do not focus on the use of the fresh food vouchers. It is expected that the participants find these activities interesting and valuable, and therefore are willing to take part without additional encouragement.

Annex 2: Theory of Change


General Assumptions:
There are links to other programmes addressing health, WASH, etc. that fight other causes of stunting
No collusion between traders and beneficiaries to buy alternative products/ exchange for cash
Local fresh foods are available yearlong
Sustained supply of fresh foods in the region
selling fresh foods is profitable (stocking and rotting issues can be managed)
Retailers willingly invest in their capacity of supplying fresh food because of future increased demand
Stunting is reduced in Amhara
accountability ceiling
PLW and children <2 adopt a healthier, more diverse diet
Beneficiaries have physical access to fresh food
Beneficiaries have financial access to fresh food
Households make proper use of nutritious food
Local markets guarantee a sustainable supply of fresh foods
Households increase their conditional food budget
Practice: improved cooking skills and dietary habits around different FF
local traders connect with producers of selected fresh food
e-voucher system is set up
participants are able to use the tools available
attitude: beneficiaries are comfortable eating taboo foods
traders have improved capacity to respond to increased demand
traders identify economic potential in FF
Knowledge: improved understanding of the nutritional aspects of FF
General Assumptions:
PSNP Programme still provides a food basket (or cash) covering their staple food needs (15 kg of grain and 4 kg of pulses)
Good nutrition in the first 1000 days of the child's life (including in the womb) is critical in improving child's health
increased income leads to consumption of fresh food
Allocation of food within the household prioritizes PLW and children <2
Cultural acceptability of fresh foods
Drivers of Change:
partnerships with NGOs and govt.
Community mobilization
Traders selection and mobilization
nutrition education
E-vouchers set up

Source: Terms of Reference (TOR) for the Impact Evaluation of WFP's Fresh Food Voucher Pilot Programme 2017-2018.

Annex 3: Endline Survey Implementation

Training of the supervisors and enumerators took place between 3 and 4 December 2018 in Mersa town in North Wollo. The training was conducted by Dr Kaleab Baye, Mrs Woinshet Tizazu and Dr Paulos Getachew Teshome from Addis Ababa University.

The training was organized in two parts. The first part was focused on the conceptual aspects of the questionnaires, module by module. Data collection and transfer protocols were also part of the training program. The enumerators were also carefully trained to take anthropometric measurements (height and weight) of children and their mothers. The second part was the field testing that took place at the end of the training, which helped to reinforce what was learnt during training as well as in detecting any remaining errors or typos in the questionnaire.

Of note is that all enumerators received the same training and the treatment status of the village/households was not revealed to the field teams. Consequently, the content of the training or the survey instrument did not vary across treatment arms. The supervisors and enumerators received financial compensation for their work.

The fieldwork took place 5-20 December 2018. In this endline survey, the teams were instructed to re-visit all 574 households that were interviewed during the baseline in December 2017. Table A1 shows the achieved sample size by study arm and by survey round. The final endline survey sample consists of 535 households, 39 households less than in the baseline. This gives an overall attrition rate of 6.8 percent, which is below the budgeted 10 percent (see Section 1.3). Compared to the control group, the attrition rates are higher among the two voucher groups.

Table A1: Sample size by treatment status; target and realized

	Total	Control	Small voucher	Large voucher
Baseline	574	140	218	216
Endline	535	135	202	198
Attrition	6.8%	3.6%	7.3%	8.3%

As a quality control measure, supervisors conducted re-visits to randomly selected households and checked that the data entered by the enumerator was accurate. If inconsistencies were detected, the enumerator was sent back to the household. Such cases were extremely rare, perhaps reflecting the high competence and motivation of the field staff.

A total of 96 traders in Mersa, Haro, Hara and Dire Roka were involved in the Focus Group Discussions (FGDs). The following questions guided the discussions:

- How do you rate the demand for fresh foods in your community?
- Which foods/food groups have the highest demand?
- How is demand affected by seasonality?

- Are there any supply side constraints
  - Food loss
  - Seasonality of supply
- How do you evaluate the price of fresh foods in the past month?
- How do you explain (causes) price changes, if any
- If any, what do you think is driving the price change?
- Who are your target clients?
  - For fruits
  - For vegetables
  - Eggs/Animal source foods
- What is the role of male partners/husbands in the market?
  - As a trader?
  - As a consumer?
- Are you selling your produce?
- If yes, how much of it represents (%) from the whole?
- If the whole food is brought on sale, ask why?
- How has the WFP's fresh Food voucher System affected the market for fresh foods in this market?
  - For fruits
  - For vegetables
  - Eggs/Animal source foods
- If demand has increased, ask how did they traders ensure sufficient supply?
- Does the voucher system work well?
- If not; what are the main problems?
- Would you say that your business improved as results of the fresh food voucher system?
- Would you say that other traders that are not part of the system have suffered?

146. FGDs were conducted with 60 caregivers of young children (18-29 months of age), both FFV and non-FFV recipients. The FGD was intended to collect qualitative information on:

- The enablers and barriers for dietary diversity in particular fresh foods consumption;
- Household dietary-related decision making,
- Identify possible misconceptions and taboos that could be targeted by the planned SBCC.
- Knowledge about the FFV program and their impressions
- Opportunities and challenges of the FFV program

147. Finally, we note that the participation to the study (both quantitative and qualitative components) was entirely voluntary and the study participants did not receive financial or any other compensation to take part in the study.

Annex 4: Assessing baseline balance

In this Annex we briefly assess the baseline balance between the different study arms. Table A2 below presents the results based on the sampling weights described in Section 1.3.[14] We see that the sample is well balanced with p-values for observed differences all above 0.05.

Table A2: Baseline balance

	(1)	(2)	(3)	t-test	t-test	t-test
	Control	Small voucher	Large voucher	Difference	Difference	Difference
Variable	**Mean/SE**	**Mean/SE**	**Mean/SE**	**(1)-(2)**	**(1)-(3)**	**(2)-(3)**
MAD	0.108	0.060	0.086	0.047	0.022	-0.025
	[0.026]	[0.017]	[0.052]			
Minimum dietary diversity	0.164	0.101	0.122	0.063*	0.042	-0.021
	[0.030]	[0.019]	[0.049]			
Minimum meal frequency	0.707	0.643	0.722	0.064	-0.014	-0.079
	[0.042]	[0.041]	[0.047]			
MDD-W	0.007	0.008	0.000	-0.001	0.007	0.008
	[0.007]	[0.005]	[0.000]			
HDDS	5.984	5.739	5.507	0.245	0.477	0.232
	[0.206]	[0.211]	[0.322]			
WFP FCS-N: Protein frequency	6.933	6.845	6.622	0.088	0.311	0.222
	[0.031]	[0.054]	[0.221]			
WFP FCS-N: Vitamin-A frequency	1.569	1.588	1.231	-0.019	0.338	0.357
	[0.304]	[0.235]	[0.154]			
WFP FCS-N: Hem iron frequency	0.342	0.334	0.335	0.008	0.007	-0.001
	[0.137]	[0.103]	[0.135]			
Caregiver's nutrition knowledge score	5.191	5.034	5.060	0.157	0.131	-0.026
	[0.514]	[0.278]	[0.360]			
Child's age in months	11.546	11.339	11.663	0.207	-0.117	-0.324
	[0.295]	[0.261]	[0.294]			
Child is boy	0.570	0.498	0.585	0.072	-0.014	-0.087*
	[0.039]	[0.028]	[0.041]			
Mother has some formal education	0.277	0.278	0.377	-0.001	-0.100	-0.099*
	[0.047]	[0.045]	[0.038]			
Household size	5.370	5.146	5.176	0.224	0.194	-0.030
	[0.169]	[0.148]	[0.290]			
Tropical Livestock Units (TLU)	2.441	2.834	2.113	-0.394	0.328	0.721
	[0.218]	[0.304]	[0.343]			

[14] We did not apply these sampling weights in the baseline report and this explains the differences between the numbers reported in the baseline and here in the endline report.

Z-score of asset index	0.105	-0.002	0.154	0.107	-0.050	-0.156
	[0.143]	[0.141]	[0.254]			
PSNP household	0.088	0.144	0.120	-0.055	-0.031	0.024
	[0.023]	[0.031]	[0.035]			
Geodetic distance (in km) to the nearest FFV market	5.775	4.193	5.278	1.582	0.497	-1.085
	[0.887]	[0.742]	[0.983]			
N	140	218	216			
Clusters	20	20	20			

*The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Standard errors are clustered at the cluster (= village or group of villages) level and reported in brackets.*

Annex 5: Evaluation matrix

Evaluation Question 1. What are the differential impacts of the programme on diet diversity for the different voucher values?						
No	Sub-questions	Measure/ indicator of progress	Main Sources of Information	Data Collection methods	Data Analysis Methods	Evidence availability / reliability
1	What is the impact of the FFV on children's (6-23 months) dietary diversity?	Minimum Diet Diversity (MDD) for children aged 6 to 23 months	Section 10 in the quantitative survey instrument. Key informant: the primary caregiver of the child.	Household Panel surveys (before and after)	Difference in Difference	3 = Strong
2	What is the impact of the FFV on children's (6-23 months) meal frequency?	Minimum Meal Frequency (MMF) for children aged 6 to 23 months	Same as above.	Household Panel surveys (before and after)	Difference in Difference	3 = Strong
3	What is the impact of the FFV on children's (6-23 months) Minimum Acceptable Diet (MAD)?	Minimum Acceptable Diet Scores (MAD) for children aged 6 to 23 months	Same as above.	Household Panel surveys (before and after)	A Difference in Difference	3 = Strong
4	What is the impact of the FFV on women's Minimum Diet Diversity?	Minimum Diet Diversity for Women (MDD-W) in the reproductive age (15-49 years)	Same as above.	Household Panel surveys (before and after)	Difference in Difference	3 = Strong
5	What is the impact of the FFV on Household Diet Diversity Score?	Household Diet Diversity Score (HDDS)	Section 7 in the quantitative survey instrument. Key informant: the primary caregiver of the child.	Household Panel surveys (before and after)	Difference in Difference	3 = Strong
6	What is the impact of the FFV on households' food consumption score?	Food Consumption Score-Nutrition	Same as above.	Household Panel surveys (before and after)	Difference in Difference	3 = Strong
Evaluation Question 2. What are the main changes in knowledge, attitude and practices of the beneficiary households regarding access and use of nutritious foods?						

No	Sub-questions	Measure/indicator of progress	Main Sources of Information	Data Collection methods	Data Analysis Methods	Evidence availability / reliability
1	What is the impact of the intervention on caregivers' knowledge of the benefits of nutritious foods?	Nutrition knowledge score (developed by the evaluators)	Section 9 in the quantitative survey instrument + FGDs with caregivers before and after the implementation. Key informant: the primary caregiver of the child.	Household Panel surveys (before and after)	Before and after comparison	2 = Fair
2	What is the impact of the intervention on caregivers' views food taboos related to fresh foods	Nutrition knowledge score related to food taboos only (developed by the evaluators)	Section 9 in the quantitative survey instrument + FGDs with caregivers before and after the implementation. Key informant: the primary caregiver of the child.	Household Panel surveys (before and after) + FGDs with caregivers before and after the implementation.	Before and after comparison	2 = Fair
3	What is the impact of the intervention on caregivers' child feeding practices	Child feeding practice score (developed by the evaluators)	Section 9 in the quantitative survey instrument + FGDs with caregivers before and after the implementation. Key informant: the primary caregiver of the child.	Household Panel surveys (before and after) + FGDs with caregivers before and after the implementation.	Before and after comparison	2 = Fair
Evaluation Question 3. Which transfer value is more cost-effective in delivering nutritional results?						

No	Sub-questions	Measure/indicator of progress	Main Sources of Information	Data Collection methods	Data Analysis Methods	Evidence availability / reliability
1	Which transfer value is more cost-effective in improving children's dietary diversity score?	(Number of food groups consumed by the child in the past 24 hours) / (per capita value of the transfer received in the past month)	Section 10 + costing data.	Household Panel surveys (before and after)	Difference in Difference	3 = Strong
2	Which transfer value is more cost-effective in improving women's dietary diversity score?	(Number of food groups consumed by the mother in the past 24 hours) / (per capita value of the transfer received in the past month)	Section 10 + costing data.	Household Panel surveys (before and after)	Difference in Difference	3 = Strong
Evaluation Question 4. What are the impacts of the project on the local markets of fresh foods?						
No	Sub-questions	Measure/indicator of progress	Main Sources of Information	Data Collection methods	Data Analysis Methods	Evidence availability / reliability
1	Has the program improved the availability of the fresh foods in the local markets?	Change in availability of fresh foods over the course of the program	Market surveys conducted by the evaluation team + qualitative data + WFP's monitoring data	Mixed methods	Descriptive analysis of the qualitative and quantitative data	1 = Weak

Annex 6: Additional food security indicators

WFP's Food Consumption Score (FCS) is a commonly used measure of food security (WFP 2008). FCS is a weighted index that combines dietary diversity and food frequency. The index is based on the household consumption of 9 food groups, with nutritionally more dense groups receiving a higher weight. The FCS ranges between 0 and 112, with lower scores indicating higher food insecurity. WFP further categorizes household diets as poor if the FCS is below 21, borderline if the score is above 21 but below 35 and acceptable if above 35. While FCS was not one of the outcome indicators, we report it here as it is closely related to the WFP's Food Consumption Score- Nutrition that is one of the outcome indicators in this evaluation.

At the baseline, the mean score in our sample was 49, with minimum value at 21 and maximum 108 (Table A3). The households in the large voucher arms had lower mean FCS than the households in the control arm ($p<0.05$). Meanwhile the difference in the mean FCS value between control arm and small voucher arm was not statistically different from zero. The mean score at the endline was 51.6 (median 49) with minimum value at 13 and maximum 100. The differences between the control arm and the two treatment arms were not statistically significant at the endline. Nor were the changes in FCS (reported in last column) between the baseline and the endline.

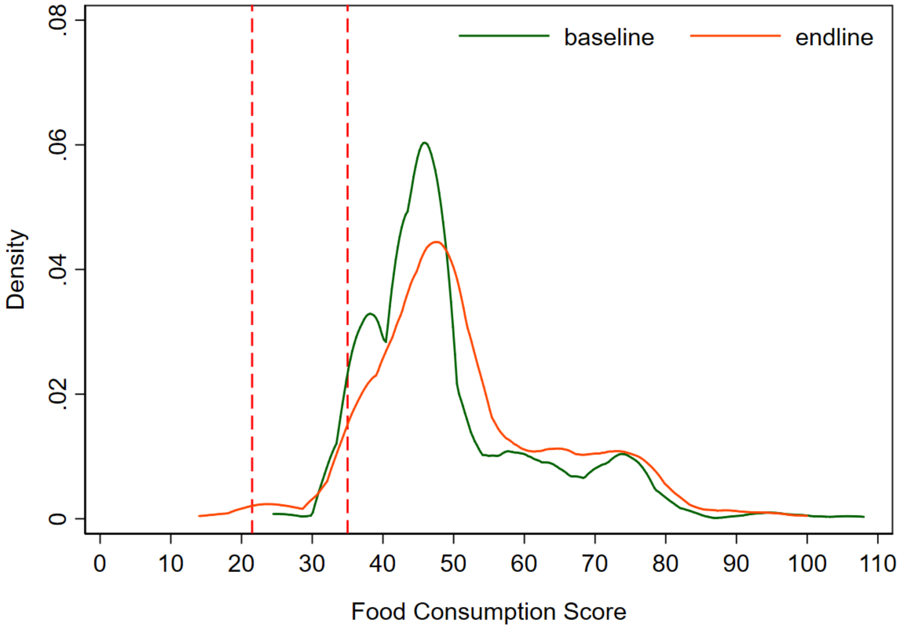
Figure A1 provides the full distribution of the FCS in both survey rounds. In both rounds, for more than 90 percent of the sampled households, the FCS is at acceptable level. For less than 10 percent of the households, the FCS score is at borderline level. Very few households (<1 %) had a score that was below 22 indicating poor FCS (Table A4). These portions are very similar that were estimated for the Amhara region by CSA and WFP using data from 2011 Welfare Monitoring Survey (CSA and WFP 2014).

Table A3: Mean Food Consumption Score (FCS), by survey round and study arm

	Baseline	**Endline**	**Difference**
Non-voucher	51.3	53.7	2.4
Small voucher	50.0	53.2	3.2
Large voucher	47.7 **	50.5	2.9
All households	**49.4**	**52.3**	**2.9**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

Figure A1: Distribution of Food Consumption Score (FCS)



Note: Kernel density. The dashed vertical lines represent the thresholds for poor (FCS<21), borderline (21.5<FCS<35) and acceptable (FCS>35) levels of food consumption.

Table A4: % of households in each FCS category, by survey round

	Baseline		Endline	
FCS Category	**N**	**%**	**N**	**%**
Poor (0-21)	0	0	4	0.75
Borderline (21.5-35)	47	8,19	36	6.73
Acceptable (>35)	527	91.81	495	92.52
Total	**574**	**100**	**535**	**100**

We also administered the Household Food Insecurity Access Scale (HFIAS) developed by Coates, Swindale, and Bilinsky (2007). Using nine different questions to explore households' perceptions on food security and their individual coping mechanisms, HFIAS has been found to provide a reliable measure of food security in different contexts (Melgar-Quinonez et al. 2006; Knueppel, Demment, and Kaiser 2010), including Ethiopia (Maes et al. 2009). In the food security module of the survey, the respondents were first asked whether they had experienced a food security issue in the last 30 days, such as a concern that their household would not have enough food. If the response was positive, then the frequency of this occurrence was ascertained. For the computation of HFIAS, a household received zero points if it reported that the event did not happen during the last 30 days, 1 if it rarely occurred (1 or 2 times), 2 if it sometimes (3 to 10 times) occurred, and 3 if it occurred often (more than 10 times). The sum of these frequency scores for the nine questions then yields a *food insecurity score* ranging between 0 and 27 with higher values indicating higher food insecurity.

While HFIAS is not one of the indicators tracked in this evaluation, we report these here to gain further insights on the food security situation in the district. Table A5 shows the mean food insecurity score in both rounds. At the baseline, mean HFIAS score in the sample was 4.9 units and the households in the small voucher arm reported higher food insecurity than the households in the control arm (p<0.05). We see that the food insecurity are marginally lower in the endline indicating small improvement in food security. About 20 percent of the households reported to be fully food secure (zero incidences of these nine food insecurity measures in the past 30 months) and this share remained the same at the endline. The remaining 80 percent of the sample reported some level of food insecurity with varying degree. The differences between the control arm and the two treatment arms were not statistically significant at the endline. However, the decrease in HFIAS between the two rounds was faster in small voucher group compared to the control group (p<0.05).

Table A5: Mean Household Food Insecurity Access Scale (HFIAS) score, by survey round and study arm

Study arm	Baseline	Endline	Difference
Non-voucher	4.2	4.2	0.0
Small voucher	5.9 **	4.5	-1.5 **
Large voucher	4.5	4.2	-0.4
All households	**4.9**	**4.3**	**-0.6**

*Note: difference to the control group tested using an weighted least squares regression where the weights were based on survey weights described in section 1.3 ; ***, **, and * indicate significance at the 1, 5, and 10 percent critical level, respectively. Standard errors were clustered at the cluster (= village or group of villages) level.*

Annex 7: Bibliography

Abay, K., and K. Hirvonen. 2017. "Does market access mitigate the impact of seasonality on child growth? Panel data evidence from North Ethiopia." *Journal of Development Studies* 53 (9):1414–1429.

Abebe, Z., G. D. Haki, and K. Baye. 2016. "Health Extension Workers' Knowledge and Knowledge-Sharing Effectiveness of Optimal Infant and Young Child Feeding Are Associated With Mothers' Knowledge and Child Stunting in Rural Ethiopia." *Food and Nutrition Bulletin* 37 (3):353 - 363.

Bachewe, F., K. Hirvonen, B. Minten, and F. Yimer. 2017. The rising costs of nutritious diets in Ethiopia ESSP Research Note 67, Addis Ababa: IFPRI.

Berhane, G., D. O. Gilligan, J. Hoddinott, N. Kumar, and A. S. Taffesse. 2014. "Can Social Protection Work in Africa? The Impact of Ethiopia's Productive Safety Net Programme." *Economic Development and Cultural Change* 63 (1):1-26.

Berhane, G., K. Hirvonen, and J. Hoddinott. 2016. The Implementation of the Productive Safety Nets Programme, 2014: Highlands Outcomes Report (2015). Addis Ababa: Ethiopia Strategy Support Program, International Food Policy Research Institute.

Berhane, G., K. Hirvonen, J. Hoddinott, S. Kim, A. S. Taffesse, K. Abay, M. Hiluf, B. Koru, T. Assefa, and F. Yimer. 2017. Evaluation of the Nutrition Sensitive Features of the Productive Safety Nets Programme IV: Baseline Survey Report. Unpublished report.

Berhane, G., K. Hirvonen, J. Hoddinott, N. Kumar, A. S. Taffesse, Y. Yohannes, M. Tefera, B. Nishan, J. Lind, R. Sabates-Wheeler, and A. Strickland. 2015. The Implementation of the Productive Safety Nets Programme in the Ethiopian Lowlands, 2014: Program Performance Report. Addis Ababa: Ethiopia Strategy Support Program, International Food Policy Research Institute.

Berhane, G., J. Hoddinott, N. Kumar, and A. Margolies. 2016. The impact of the Productive Safety Net Programme on schooling, child labour and the nutritional status of children in Ethiopia. 3ie Grantee Final Report. , New Delhi: International Initiative for Impact Evaluation (3ie)

Breitenstein, S. M., D. Gross, C. A. Garvey, C. Hill, L. Fogg, and B. Resnick. 2010. "Implementation fidelity in community-based interventions." *Research in nursing & health* 33 (2):164-173.

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14 (5):365.

Coates, J., A. Swindale, and P. Bilinsky. 2007. Household Food Insecurity Access Scale (HFIAS) for measurement of food access: indicator guide. Washington, DC: Food and Nutrition Technical Assistance Project, Academy for Educational Development.

CSA. 2010. Population and Housing Census Report- Amhara Region. Addis Ababa: Central Statistical Agency (CSA) of Ethiopia.

CSA, and ICF. 2016. Ethiopia Demographic and Health Survey 2016. Addis Ababa, Ethiopia, and Rockville, Maryland, USA: Central Statistical Agency (CSA) of Ethiopia and ICF.

Dercon, S., and P. Krishnan. 2000. "In sickness and in health: Risk sharing within households in rural Ethiopia." *Journal of Political Economy* 108 (4):688-727.

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

Durlak, J. A., and E. P. DuPre. 2008. "Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation." *American journal of community psychology* 41 (3-4):327.

FAO, and FHI 360. 2016. Minimum Dietary Diversity for Women: A Guide to Measurement. Rome: Food and Agriculture Organization (FAO) of the United Nations and USAID's Food and Nutrition Technical Assistance III Project (FANTA), managed by FHI 360.

FDRE. 2017. Growth and Transformation Plan II (GTP II) (2015/16-2019/20). Addis Ababa: Federal Democratic Republic of Ethiopia (FDRE); The National Planning Commission.

GFDRE. 2014a. Productive Safety Net Programme 4: Design Document. Addis Ababa: Government of the Federal Democratic Republic of Ethiopia (GFDRE).

———. 2014b. Productive Safety Net Programme: Programme Implementation Manual. Addis Ababa: Government of the Federal Democratic Republic of Ethiopia (GFDRE).

———. 2016. National Nutrition Programme 2016-2020. Addis Ababa: Government of the Federal Democratic Republic of Ethiopia (GFDRE).

Headey, D., K. Hirvonen, J. Hoddinott, and D. Stifel. 2018. Rural food markets and child nutrition.

Hirvonen, K., A. Wolle, and B. Minten. 2018. Affordability of fruits and vegetables in Ethiopia. In *IFPRI-ESSP Research Note*. Washington D.C.: IFPRI-ESSP.

Kim, S. S., D. Ali, A. Kennedy, R. Tesfaye, A. W. Tadesse, T. H. Abrha, R. Rawat, and P. Menon. 2015. "Assessing implementation fidelity of a community-based infant and young child feeding intervention in Ethiopia identifies delivery challenges that limit reach to communities: a mixed-method process evaluation study." *BMC public health* 15 (1):316.

Kim, S. S., R. Rawat, E. M. Mwangi, R. Tesfaye, Y. Abebe, J. Baker, E. A. Frongillo, M. T. Ruel, and P. Menon. 2016. "Exposure to large-scale social and behavior change communication interventions is associated with improvements in infant and young child feeding practices in Ethiopia." *PLoS ONE* 11 (10):e0164800.

Knueppel, D., M. Demment, and L. Kaiser. 2010. "Validation of the household food insecurity access scale in rural Tanzania." *Public Health Nutrition* 13 (3):360-367.

Maes, K. C., C. Hadley, F. Tesfaye, S. Shifferaw, and Y. A. Tesfaye. 2009. "Food insecurity among volunteer AIDS caregivers in Addis Ababa, Ethiopia was highly prevalent but buffered from the 2008 food crisis." *The Journal of Nutrition* 139 (9):1758-1764.

Melgar-Quinonez, H. R., A. C. Zubieta, B. MkNelly, A. Nteziyaremye, M. F. D. Gerardo, and C. Dunford. 2006. "Household food insecurity and food expenditure in Bolivia, Burkina Faso, and the Philippines." *The Journal of Nutrition* 136 (5):1431S-1437S.

MoANR, and MoLF. 2016. Nutrition Sensitive Agriculture Strategy. Addis Ababa, Ethiopia: Ministry of Agriculture and Natural Resource (MoANR) and Ministry of Livestock and Fisheries (MoLF).

NDRMC. 2018. Ethiopia: Humanitarian and Disaster Resilience Plan 2018, Mid-Year Review

Addis Ababa: Joint Government and Humanitarian Partners' Document. National Disaster Risk Management Commission (NDRMC).

Stifel, D., and B. Minten. 2017. "Market Access, Welfare, and Nutrition: Evidence from Ethiopia." *World Development* 90:229-241.

Swindale, A., and P. Bilinsky. 2006. "Household dietary diversity score (HDDS) for measurement of household food access: indicator guide." *Washington, DC: Food and Nutrition Technical Assistance Project, Academy for Educational Development*.

UNDP. 2016. *2016 Human Development Report: Human Development for Everyone*. New York, USA: United Nations Development Programme (UNDP).

WFP. 2008. Food consumption analysis: Calculation and use of the food consumption score in food security analysis. Rome: World Food Programme (WFP), Vulnerability Analysis and Mapping Branch (ODAV).

———. 2015. Food Consumption Score Nutritional Quality Analysis Guidelines (FCS-N). Rome: United Nations World Food Programme (WFP), Food security analysis (VAM).

WHO. 2008. Indicators for assessing infant and young child feeding practices: part 1: definitions: conclusions of a consensus meeting held 6-8 November 2007 in Washington DC, USA. Geneva: World Health Organization (WHO).

———. 2010. *Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies*. Geneva: World Health Organization (WHO).

Zerfu, T. A., M. Umeta, and K. Baye. 2016. "Dietary habits, food taboos, and perceptions towards weight gain during pregnancy in Arsi, rural central Ethiopia: a qualitative cross-sectional study." *Journal of Health, Population and Nutrition* 35 (22):1-7.

List of Acronyms

AAU	Addis Ababa University
CO	Country Office
DEQAS	Decentralized Evaluation Quality Assurance System
DHS	Demographic and Health Survey
EB	Executive Board
EC	Evaluation Committee
ERG	Evaluation Reference Group
ESSP	Ethiopia Strategy Support Program
ET	Evaluation team
FCS	Food Consumption Score
FGDs	Focus Group Discussions
FFV	Fresh Food Voucher
FTMA	Farm to Market Alliance
GEEW	Gender equality and women's empowerment
HEW	Health extension worker
HDA	Health development army
HH	Household
HDDS	Household dietary diversity score
HQ	Headquarters
IFPRI	International Food Policy Research Institute
IYCF	infant and young child feeding practices
KfW	Kreditanstaltfür Wiederaufbau
M&E	Monitoring and Evaluation
MAD	minimum acceptable diet (children aged 6 to 23 months)
MDD	minimum dietary diversity (children aged 6 to 23 months)
MDD-W	minimum dietary diversity for women of reproductive age
MMF	minimum meal frequency (children aged 6 to 23 months)
NGO	Non-Governmental Organization
OEV	Office of Evaluation
PLW	pregnant and lactating women
PSNP	Productive Safety Net Program
QS	Quality Support
RB	Regional Bureau
SBCC	Social Behaviour Change Communication
SOP	Standard Operating Procedure
TOC	Theory of Change
TOR	Terms of Reference
UNCT	United Nations Country Team
UNDSS	UN Department of Safety & Security
UNEG	United Nations Evaluation Group
VAM	Vulnerability and Analysis Mapping
WFP	World Food Programme
WFP ETHCO	World Food Programme - Ethiopia Country Office
WHO	World Health Organization

Addis Ababa, May 2019

WFP Ethiopia Country Office

http://www1.wfp.org/countries/ethiopia

World Food Programme