Western Highlands Integrated Program Evaluation

Addendum to the Baseline Report: Sampling Considerations

July 2018

USAID
FROM THE AMERICAN PEOPLE

MEASURE Evaluation

Western Highlands Integrated Program Evaluation

Addendum to the Baseline Report: Sampling Considerations

Roberto Molina-Cruz, MSc, consultant
Tory M. Taylor, MPH, MEASURE Evaluation
Gustavo Angeles, PhD, MEASURE Evaluation

Cover: A young woman prepares tortillas for sale at a market in the town of Chichicastenango, Guatemala, department of El Quiché. Photo: ©2013 Tory M. Taylor, MEASURE Evaluation.

July 2018

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 W. Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: +1 919-445-9350 | measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. TRE-18-011

ISBN: 978-1-64232-050-3

USAID
FROM THE AMERICAN PEOPLE

MEASURE
Evaluation

CONTENTS

ABBREVIATIONS

INE	Instituto Nacional de Estadística (National Statistics Institute)
RVCP	Rural Value Chains Project
WHIP	Western Highlands Integrated Program
ZOI	zone of influence

INTRODUCTION

This document is an addendum to the *Monitoring and Evaluation Survey for the Western Highlands Integrated Program: Baseline 2013* report released in 2014 (Angeles, Hidalgo, Molina-Cruz, Taylor, Urquieta-Salomón, Calderón…Romero, 2014). It contains detailed additional technical information on the sampling procedures used in the baseline household survey, including the development of the sampling frame for domains specific to the Rural Value Chains Project (RVCP); the initial selection procedures for primary and secondary sampling units (census tracts and households, respectively); and procedures for sampling unit replacements. The content is organized in three sections: 1) unit identification and selection processes for sample domains 1 and 2, which relied heavily on beneficiary lists obtained from program partners and subsequent household mapping and verification activities; 2) the necessary deviations from probability-based approaches in census tract selection and possible analytic implications; and 3) the replacement of households in the sample and likelihood of affiliated bias in the baseline results.

A complete description of the Western Highlands Integrated Program (WHIP), the design of the program evaluation, and the baseline survey results are provided elsewhere (Angeles, et al., 2014; Taylor, 2014). In brief, the evaluation intended to generate both performance and impact estimates over time across a range of health, nutrition, and economic indicators reflecting program activities and goals. The impact evaluation component uses a quasi-experimental difference-in-differences design with propensity score-matched comparison groups, as described below. Sampling was designed to capture differences between subgroups in the integrated program's zone of influence (ZOI), especially based on exposure to the program's RVCP. In 2013, baseline surveys were conducted with members of 6,301 households in 226 census tracts representing 54 municipalities. Unit selection procedures and sample sizes were determined separately for each of five domains:

1. RVCP participant households, who were also exposed to the health program (RVCP direct beneficiaries)

2. RVCP non-participant households located in the same census tracts as the RVCP participants, who were also exposed to the health program (RVCP indirect beneficiaries)

3. Households located outside the census tracts where RVCP participants resided, who were also exposed to the health program (health only)

4. Comparison households in the census tracts similar to those in domains 1 and 2

5. Comparison households in the census tracts similar to those in domain 3

Taken together, results from domains 1 through 3 reflect the situation in the ZOI. By comparing changes over time among the domains, the following primary and secondary research questions can be answered:

- What are the changes in key indicators at the population level in the ZOI?

- What are the effects of the WHIP on key indicators at the population level in the ZOI?

- What changes are present in key outcomes at the population level in the three groups constituting the ZOI?

- What has been the impact of the integrated program on key results at the population level among RVCP direct beneficiaries and RVCP indirect beneficiaries?

- What has been the impact of the health and nutrition program, acting without the RVCP, on key outcomes at the population level in the "health only" domain?

- Is the integrated program more effective than the health and nutrition program alone in improving key outcomes at the population level?

- Does the RVCP have indirect effects on the non-member households located in RVCP areas? If so, what are these effects?

RVCP SAMPLE FRAME AND UNIT SELECTION

A standard two-stage sampling procedure was planned for domains 1 and 2. The first stage involved the selection of census tracts that included RCVP producers' association member households, with selection probability proportionate to the number of member households in the tract. The second stage involved the random selection of a fixed number of member and non-member households in each selected tract. Membership information obtained from RVCP-supported producers' associations was planned to be used to identify the set of census tracts where members resided. This process also intended to use census tract maps provided by the Instituto Nacional de Estadística (INE, or National Statistics Institute) and generated following the 2002 national census. The maps show the boundaries, landforms, major structures, and other reference points in each tract. Any given tract contains approximately 200 to 350 housing units located within a reasonable walking distance from one another.

An updated map for each tract selected for the sample was prepared using a map of the same area from a more recent survey and aerial photographs of the area obtained from the INE. Cartographic experts were dispatched to confirm the physical boundaries of the tracts and update existing structures and landforms on the maps. FC-01 forms were used to identify the members of households included in the tracts. The FC-01 is a standard INE-issued household registry form used for the national census and other official surveys in Guatemala showing the number of household members, number of male habitual residents, number of female habitual residents, the name of the head of household, the household's geographic location, and the physical characteristics of the residence and the surrounding area.

Households containing one or more RVCP members, as suggested by the information obtained from the RVCP associations, constituted the initial sample frame for domain 1. All other households in these tracts constituted the sample frame for domain 2. By noting the total number of member and non-member households in each tract, we planned to select a predetermined number of households, by type, in accordance with the calculated sample sizes (20 households in tracts in domains 1 and 2, and 30 households in tracts in domain 3). In each selected census tract, the target number of households, by type, would be sampled with equal probability (simple random sampling). Selection would use a systematic procedure, e.g., by identifying a starting household and a fixed sample interval with which the other households were selected. Households selected for the sample were marked on copies of the tract maps.

Although every attempt was made to obtain complete and accurate data about members' home addresses from the RVCP associations and to accurately match these addresses to the maps and data tables obtained from the INE, in a subset of cases, we were unable to identify the specific tract where an association member resided. In these cases, we identified the "census area" instead, defined as the smallest contiguous set of tracts believed to contain the dwelling(s) in question. The sample frame for domains 1 and 2 was therefore developed using two different primary sampling unit types:

(a) Census tracts where one or more RVCP members resided

(b) Census areas encompassing two or more tracts, where one or more RVCP members were believed to reside

Size estimates were generated for each of the 213 primary sampling units based on the estimated number of RVCP member households. Census tracts and areas were classified into three strata: the first containing those with at least 100 member households (n=25); the second containing those with between 20 and 99 member households inclusive (n=87); and the third containing those with fewer than 20 member households (n=101). All 25 units in the first stratum were selected in stage 1, and 35 units were selected from each of the second and third strata using probability proportionate to size. This resulted in an initial sample of 95 primary sample units for domains 1 and 2. For the selected census areas, we inquired again at the RVCP associations' offices to more precisely determine the location of their members' homes, if possible. One or two tracts in the census area were then selected with probability proportionate to size, with the size measure being the estimated number of RVCP households, and household selection in the tracts proceeded as previously described.

Census Tract Replacement

During the development of the domain 1 sample frame, the research team reviewed lists of association members provided by the associations. The information on the lists was limited to members' names and household locations, details that were frequently imprecise. To compensate, additional information obtained from the INE was applied to these efforts; however, this information largely originated from the 2002 census and, as such, also proved to be outdated. The research teams therefore initiated fieldwork aware of the potential deficiencies in the sample frame for this domain. Indeed, 12 of the 95 primary sampling units initially selected for the domain 1 sample (some tracts and some census areas) were found not to contain a household that included an association member. On the first discovery at the start of fieldwork, and to avoid a significant reduction in the sample size for domains 1 and 2 due to inaccuracies in the sample frame, the team developed a substitution procedure for replacing these ineligible sampling units with qualifying ones. Primary sampling unit replacement of this type is common in the Demographic and Health Surveys and other household-based studies, typically because of natural disasters or other safety concerns in a location originally designated as part of the sample (ICF International, 2012). In addition to preserving the accuracy of the sample frame and the estimated unit selection probabilities, the procedure protects against household losses leading to a reduced sample size.

Primary sampling unit replacement was not conducted at random. From among the 52 eligible tracts not originally selected as part of the sample in domains 1 and 2, we selected 38 deemed most likely to have RVCP association member residents, based on the quality of information provided by the associations. From those 38, the fieldwork coordinators identified 20 whose addition to the sample would minimally disrupt established plans for fieldwork. Despite being purposive, this selection was presumed not to correlate with the main variables under study because fieldwork planning only accounted for general characteristics of the census tracts and was principally defined by logistical considerations. Ultimately, 12 of the 20 census tracts were used. The need for substitutions was identified early in the fieldwork process, during the household listing and verification efforts that preceded the arrival of interview teams in a given area. We assigned probabilities of selection to the replacement tracts according to their probability of not being included in the initial sample, and an equal probability of selection in the additional (replacement) sample.

Household Replacement

The team's cartographers conducted a multistep household verification process, including the identification of structures containing households, enumerating these households, and obtaining information about them for the FC-01. This information was preferentially obtained from a household member, but in cases where a structure was unoccupied, neighbors were consulted. If no one was present in or near a structure that may have contained a household, it was difficult to accurately determine whether the structure was an unoccupied household or a non-household. Some households may also have been abandoned after the enumeration and verification but before the arrival of interviewers. The cartographer may therefore have concluded that a structure represented a household when, at the time of the interview, it did not and vice versa. During fieldwork, the first type of error could be identified and corrected, but the second type could not. Hence, the number of eligible households mistakenly omitted from the sample frame is unknown.

In addition to non-households originally classified as households, some domain 1 households were erroneously classified as containing an RVCP member. In these cases, we used a replacement process, selecting the nearest eligible household according to the map and the corresponding list of households in the tract. Table 1 summarizes the replacements, by survey domain.

Table 1. Household replacements, by survey domain

Domain	Tracts	Households	Households replaced N (% of total)
1	89	1,264	43 (3.4%)
2	89	1,746	167 (9.6%)
3	34	997	114 (11.4%)
Subtotal (ZOI)	212	4,007	324 (8.1%)
4	74	1,438	72 (5.0%)
5	29	856	37 (4.3%)
Subtotal	103	2,294	109 (4.8%)
Total	315	6,301	433 (6.9%)

The fact that there were households included in the frame by error (and therefore, outside of the target population) is not in and of itself a problem. Any potential issue lay in the method of selection of the replacement households. A problem of endogenous sample selection could arise if the "nearest neighbor" rule differed in its implications for representativeness from random selection sufficiently such that the selected neighbor households represented a type of household within the population with significantly different average outcomes than others in the census tract. For this to happen, it would have been the case that the misclassified households were included in the frame for reasons related to an outcome of interest (stunting or poverty) and that there was a significant correlation between those misclassified households and their neighbors. A pertinent example would be a city neighborhood where a high density of abandoned

houses indicates low socioeconomic status for the neighborhood overall, which differs from the socioeconomic status of the rest of the city.

Although we cannot definitively rule out the possibility of this situation, the evaluation team considers it very unlikely in the context of the WHIP baseline survey tracts, given that the tracts themselves are not large and so variation in the conditions of the households in the survey tracts. Moreover, for this to yield bias at the level at which estimates were reported, two things would have to be true. First, the deviation from representativeness by these "nearest neighbor" households would have to be pronounced. Second, it would have to have been a widespread problem. The assessment of this as unlikely is based on the investigation of the sources of the errors in the sample frame, a review of the procedures implemented in practice for the "nearest neighbor" rule, and the evaluation team's knowledge of the variation of household conditions in the survey tracts in the Western Highlands.

Because the weights assigned to the census tracts, households, and individuals in them are proportionate to the combined number of households and non-households in the sample as determined during the cartography updates, they reflect overestimates in tracts containing non-households. To the extent that having non-households included in the sample for a given tract might be associated with other characteristics of interest, the use of the weights could introduce bias into indicator/results estimates at the domain level.

To account for this possibility, we adjusted the household, women's, and children's case weights as follows (based on estimates of the proportion of non-households in the tract):

$$w' = w \cdot m / (m + c)$$

where:

w' = Adjusted household weight

w = Original household weight

m = Number of households interviewed in the tract

c = Number of households replaced

Adjustment to Baseline Weights

This document presents the difficulties encountered during the development of the baseline sample frame for domains 1 and 2, especially the identification and geo-locating of households where some RVCP association members resided. Because of the difficulties, some census tracts were initially selected for these domains, but during the household listing verification and mapping, they were later determined not to contain an association member in residence, and were therefore not eligible for inclusion in the two domains.

To avoid a significant reduction in the sample size for domains 1 and 2, we selected 20 additional tracts using the procedure described in this addendum. Of these 20 tracts, 12 were substituted for previously selected tracts in the sample (which had been discarded due to ineligibility).

The baseline study report describes the selection probabilities assigned to tracts included in the sample for two groups, the initial sample and the "additional" sample comprised of substitutions. The selection probabilities can be expressed as described below, where p2 is the probability that a census tract not selected for the initial sample is selected in the additional sample:

Probability	Sample
p1	Initial
(1 − p1) p2	Additional

After these calculations were made, it was brought to our attention[1] that the probabilities likely required revision because they should have been calculated considering the initial and additional sampling procedures as a single procedural step. The revised selection probability of each tract in the sample (initial or additional) is as follows:

$$p1 + (1 - p1)\, p2$$

This revision to the selection probabilities for the primary sampling units caused us to revisit the calculation of sampling weights for the units in the study. Because all weights were calculated multiplicatively, with the first factor corresponding to the weight assigned to the selected tract, re-calculation of the weights to generate corrected values was conducted using the following (single) factor applied initially to assign a new weight to the selected tract:

FACTOR1 = PSC5 / PSC6

where the following values remain as described in the baseline results report:

PSC5 = PSC1 if the census tract was selected in the initial sample.
= PSC2 if the census tract was selected in the additional sample

PSC6 = PSC1 + PSC2

At the census tract level, the FACTOR1 values are contained in the file named "WHIP baseline weights adjustment factors.xlsx" (available here: https://www.measureevaluation.org/resources/files/TRE-18-011-Excel.xlsx/at_download/file). Note that this value is equal to 1 for all sectors to which the adjustment does not apply, that is, those in domains 3, 4, and 5. There are only 42 sectors (18.6 percent of the 226 sectors in the sample) with an adjustment factor different than 1. There are 184 sectors with a value equal to 1.

Technical note on use of the adjustment factors in the file adjustment_factors.xlsx: Users should merge the sector-level adjustment factors, presented in column FACTOR1, to the 2013 baseline data files using HPAQUETE as the merging key variable, and then proceed to obtain the adjusted weights by multiplying the existing weights by FACTOR1. For example, in the household data file "hogar1.dta," the adjusted household weights are obtained by:

Adjusted_pesohogar = pesohogar * FACTOR1

Application of these adjustment factors had only a minimal effect on the baseline ZOI estimates. For example:

- For stunting children under 5, the values are:
  - Previous weights: 67.40%, confidence interval [62.69, 72.10]
  - With adjusted weights: 67.47%, confidence interval [62.38, 72.56]

- For food security, households with moderate or severe hunger:
  - Previous weights: 13.7%, confidence interval [11.04, 16.43]
  - With adjusted weights: 13.6%, confidence interval [10.76, 16.53]
- For poverty 1.25:
  - Previous weights: 5.9%, confidence interval [4.01, 7.75]
  - With adjusted weights: 5.9%, confidence interval [3.90, 7.91]

CONCLUSIONS

The 2013 baseline survey for the combined performance and impact evaluation of the WHIP in Guatemala used a complex multistage sampling procedure designed to accommodate several analytic priorities. In the two domains intended to distinguish RVCP association member households and other households in the same census tracts, sampling involved an iterative process, including working with association representatives to develop membership lists, mapping these lists to updated census tract maps, selecting both member and non-member households for the sample, and confirming the accuracy of the sampled households' characterization as containing an RVCP member or not. Limited geographical accuracy of the information obtained about the association members and other difficulties encountered in identifying households during the household listing process across survey domains led to both census tract- and household-level substitutions designed to reduce error and preserve statistical power. A close practical and theoretical examination of the replacement processes and analytical strategies applied in response to these challenges leads us to conclude that the potential for serious bias in the baseline results consequent to these procedures is minimal.

REFERENCES

Angeles, G., Hidalgo, E., Molina-Cruz, R., Taylor, T., Urquieta-Salomón, J., Calderón, C., Fernández, J.C., Hidalgo, M., Brugh, K., & Romero, M. (2014). *Monitoring and evaluation survey for the Western Highlands Integrated Program: Baseline 2013* (TR-14-100). Chapel Hill, NC, USA: MEASURE Evaluation, University of North Carolina. Retrieved from http://www.cpc.unc.edu/measure/publications/tr-14-100.

ICF International. (2012). *Demographic and health survey sampling and household listing manual*. Calverton, MD, USA: ICF International. Retrieved from https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

Taylor, T. M. (2014). *The Western Highlands Integrated Program (WHIP) evaluation baseline survey: A case study in evaluation practice* (SR-14-106). Chapel Hill, NC, USA: MEASURE Evaluation, University of North Carolina. Retrieved from http://www.cpc.unc.edu/measure/publications/sr-14-106.

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 W. Franklin Street, Suite 330
Chapel Hill, NC 27516 USA
Phone: +1 919-445-9350 | measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. TRE-18-011

ISBN: 978-1-64232-050-3

USAID
FROM THE AMERICAN PEOPLE

MEASURE
Evaluation