CATHOLIC RELIEF SERVICES

CRS Samples

A GUIDE ON CALCULATING QUANTITATIVE SAMPLE SIZES

Stacy Prieto, PhD, wrote this guide. It draws heavily from *Going beyond simple sample size calculations: a practitioner's guide* by Brendon McConnell and Marcos Vera-Hernández, making that guide more accessible to CRS staff and applicable to the CRS operating context.

Cover photo by Katie Price

Catholic Relief Services | 228 W. Lexington Street, Baltimore, MD 21201, USA | crs.org | crsespanol.org
For more information, contact stacy.prieto@crs.org.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The author would like to thank Benjamin Allen, James Campbell, Tony Castleman, Heather Dolphin, John Hembling, and TD Jose of CRS, Stephanie Martin of TANGO International, Eric Gerber of Northeastern University, and Tim Smith of the United States Census Bureau for their contributions to this guide's content.

This guide is dedicated to CRS Monitoring, Evaluation, Accountability, and Learning (MEAL) staff who work tirelessly to improve their knowledge and capacity to implement high-quality global MEAL practices.

Abbreviations

BHA	USAID Bureau for Humanitarian Assistance
BMI	Body Mass Index
CI	Confidence Interval
FPC	Finite Population Correction
FTF	USAID Feed the Future Initiative
HH	Household
ICC	Intracluster (or Intraclass) Correlation Coefficient
IPTT	Indicator Performance Tracking Table
IYCF	Infant and Young Child Feeding
MAD	Minimum Acceptable Diet
MDE	Minimum Detectable Effect
MEAL	Monitoring, Evaluation, Accountability, and Learning
MIRA	Measuring Indicators for Resilience Analysis
NGO	Non-Governmental Organization
PPS	Probability Proportional to Size
SD	Standard Deviation
SE	Standard Error
SILC	Savings and Internal Lending Communities
USAID	United States Agency for International Development
USDA	United States Department of Agriculture

Introduction

CRS and the larger international development community have become increasingly rigorous in the quantitative evaluation of development projects. CRS, in its internal Monitoring, Evaluation, Accountability, and Learning (MEAL) Policies and Procedures, requires baseline studies and final evaluations for all projects greater than $1 million in value (Catholic Relief Services 2023). After collecting final evaluation data, indicator performance tracking table (IPTT) values for outcome-level indicators should be compared to their respective baseline values to know if the means[1] or proportions of indicators differ significantly across the two time periods.

Detecting a statistically significant difference between two means or proportions requires 1) That a difference truly exists and 2) That the representative sample from which the means or proportions are calculated is sufficiently large.

In addition, data for many annual performance monitoring indicators are collected from a sample of project beneficiaries. That sample must be sufficient to reasonably estimate the indicator value among all project beneficiaries.

This guide summarizes the key considerations for determining sample sizes, with examples specific to CRS's work. It is intended to fill a gap in agency knowledge around sample size calculations, reduce confusion around which equations are appropriate for use in which contexts, and provide a consistent agency-wide reference point.

Objective

This guide should be used during the project design stage to allocate sufficient resources in the project budget for data collection and to revise data collection needs during project start-up and implementation. It includes sample size equations that calculate a Minimum Detectable Effect (MDE) size since CRS projects often seek to detect a significant change between two points (baseline and final).

The audience for this guide is proposal MEAL leads, project MEAL coordinators, country program MEAL managers and MEAL technical advisors.

Annex 6 discusses donor sample size guides and highlights some key differences between them and this guide.

The intended audience for this guide is proposal MEAL leads (for MEAL budget preparation), project MEAL coordinators (to inform data collection during start-up and implementation), and country program MEAL managers and MEAL technical advisors (to aid in their support of the above).

When to use this guide

Project staff should calculate a sample size for each sample frame or respondent type they will survey using quantitative methods. For example, suppose a project will survey farmers to track one indicator, producer groups to track another, and children under five years old to track another. They should identify the appropriate equation and calculate each sample size.

Suppose data for multiple indicators will be collected from one sample frame. In that case, the best practice is to calculate the sample size needed for each indicator and then choose the

[1] This guide uses the term "mean" throughout because it is the term used by mathematicians and statisticians.

If data for multiple indicators will be collected from one sample frame, calculate the sample size needed for each indicator, and then choose the largest size calculated.

largest calculated size within reason. If the largest sample size would be expensive to collect, focus on 1) the most important, typically outcome-level indicators or 2) donor standard indicators. Consult a MEAL technical advisor as well, if needed.

This guide can also estimate the necessary sample size for output-level indicators. However, projects typically collect data on output-level indicators via distribution or training records (routine monitoring), not a representative sample of project beneficiaries.

How This Guide is Organized

This guide provides a basic decision tree for determining the appropriate equation for calculating sample sizes. It then has a mandatory section on considerations that may increase the required sample size to ensure the final sample size is adequate. This is followed by an optional section on factors that may decrease the sample size. It is recommended only to use the latter section if the previously determined sample size is too expensive, and it is necessary to reduce it. The guide includes equations that measure change between two time periods. It also includes equations for conducting a simple assessment (not detecting a change over time or among groups).

Although this document only intends to guide sample size calculations, there is a brief section on sample size myths and how sample design affects data analysis.

The guide has six annexes: 1) intracluster correlation coefficients; 2) an Excel spreadsheet and 3) example R code for more complex computations; 4) boilerplate language for describing sample size calculations in formal written documents (e.g., donor reports, project proposals, etc.); 5) a quick reference guide; and 6) a review of other key sample size guides.

Information Needed Before Using this Guide

This guide assumes a few steps have occurred before using this guide.

1) Project teams have decided if they will make statistical comparisons between groups with the results. For example, comparing project beneficiaries (treatment group) to a control group, comparing baseline to endline values, comparing men to women, comparing age groups, etc.

2) The project's indicator performance tracking table (IPTT) with estimated baseline and target values has already been developed. Usually, the IPTT is estimated at the project design stage, with a desk review determining baseline value estimates from other similar projects or studies. These should be more recent studies covering the same or similar geographic areas, seasons, and climate conditions (e.g., drought/ non-drought).

After reviewing who should use this guide and when and gathering the necessary information, project teams can review the sample size decision tree (Section 1) to determine the required equation for each indicator.

Registering and delivering multi-purpose cash assistance in Brazil. [Felippe Thomaz]

1. Sample size decision tree

Figure 1, the Sample Size Decision Tree, represents the key question that needs to be answered for each indicator: "Which sample size equation should project teams use?" The decision tree guides users to the appropriate equation (there are eight) based on whether they will statistically compare collected data between groups, collect data for a continuous or binary indicator, and cluster the sample during data collection.

Dots (A) and (B) in Figure 1 lead to additional information to aid in decision making.

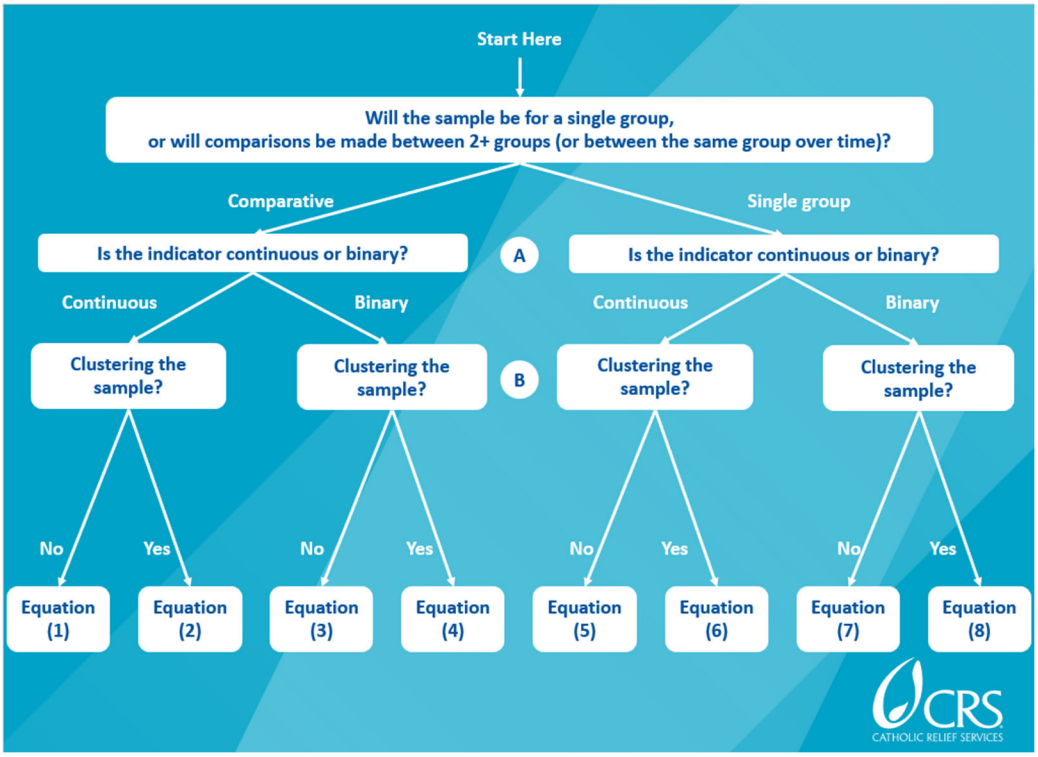


FIGURE 1. SAMPLE SIZE DECISION TREE

Dots (A) and (B) in Figure 1 provide additional information to aid decision-making. Regarding (A), continuous and binary indicators follow different probability distributions; hence, they require different sample size equations:

(A) Continuous indicators collect data with many possible values. A few examples:

- Mean value of annual sales of farms and firms[2]
- Mean number of hectares under improved management practices or technologies
- Mean yield of targeted agricultural commodities
- Mean volume of commodities sold by farms and firms

[2] Note that when using a sample to report values for standard indicators such as "Value of Sales, Number of Hectares under Improved Management Practices, Number of Individuals Using Improved Management Practices, etc.", these indicators require extrapolating from the mean value of a sample. Therefore, calculate the necessary sample size for the mean value of these indicators.

- Mean classroom attendance rate
- Mean number of food groups consumed by a household
- Index of social capital at the household level

Binary indicators collect data as a yes/ no response. A few examples:

- Percentage of individuals or organizations using an improved practice
- Percentage of students who can read
- Percent of children under 5 who are stunted
- Percentage of individuals accessing agriculture-related financing

(B) Clustering means randomly selecting a village, school, producer association, etc., as the cluster and then randomly selecting people within that cluster for the sample. Simple random sampling, by contrast, means having a list of all people in all villages and randomly selecting people from this list without first selecting a cluster.

Often, samples are clustered to save money when survey respondents are spread over a wide geographic area. Clustering the sample avoids the need to visit every village, which might be more expensive.[3]

Note that, with clustering, the sample size equations indicate the number of clusters to survey and the number of individuals within each cluster. Keeping the number of individuals per cluster as similar as possible is essential. For example, if schools in the target area are small and do not always have 15 students meeting the selection criteria, then do not use 15 individuals/ cluster in the sample size equations.[4]

[3] Clustering, while reducing data collection costs, does complicate analysis. See Section 6.2 for additional details.

[4] Evaluation designs used by development practitioners frequently use probability proportional to size (PPS) sampling methods to increase the likelihood that clusters with larger populations will be included in the sample. Please see Section 6.1.2 on data analysis for the limitations of the PPS methodology.

2. The eight main equations

The eight sample size equations are explained in Section 2 and assume that a representative sample is randomly[5] selected from the population.

2.1 Multiple study groups

All eight equations provide the sample size needed for each comparison group.

Before diving into the equations, it is essential to note that all eight equations provide the sample size needed for each comparison group. Examples of comparison groups are baseline and final or treatment and control.[6] If data is collected at baseline for comparison to final evaluation data, the final number provided should be applied in each instance. So, if the sample size is 100, 100 observations are needed at baseline and 100 at final—or 100 observations in the treatment group and 100 in the control group. Additionally, if results are analyzed among strata,[7] such as testing for statistical differences between geographic areas or gender, the final number should be applied to each stratum.

2.2 The Equations

This section will walk users through equations (1-8), explain each of the terms in each equation, and provide one practical example for each equation. Appendix 2 is an Excel spreadsheet with equations (1-8) already programmed to aid in calculations.

2.2.1 Comparisons between groups

The reference for the first four equations is McConnell and Vera-Hernández (2015). This reference is a good choice for the CRS context because 1) It provides the exact equations needed, rather than referencing packaged commands in statistical software; 2) It mathematically derives the sample size equations and/ or provides references for the derivations; 3) It provides minimum detectable effect (MDE) equations for both binary and continuous indicators.

Equations (1-2) are the same as those used in the Feed the Future guide for third-party evaluators (Stukel 2018a), whereas equations (3-4) differ somewhat. See Appendix 6 for further comparison of this guide to the Feed the Future guide.

Note that if making comparisons between groups:

- For a continuous indicator collected from a non-clustered sample, use equation (1)
- For a continuous indicator from a clustered sample, use equation (2)
- For a binary indicator from a non-clustered sample, use equation (3)
- For a binary indicator from a clustered sample, use equation (4)

See Section 1 if you are unclear on the terms “continuous,” “binary,” or “cluster.”

[5] Non-random samples and their associated bias are addressed in Section 6.1.1.

[6] Note that a future version of this guide may reference different, likely more efficient equations for determining the total sample size and allocation between 2 or more treatment groups (Duflo, Glennerster, and Kremer 2007).

[7] See Section 4.3 when stratifying the sample for organizational purposes (not to analyze differences between strata). In this case, it is not necessary to increase the sample size.

2.2.1.1 Equation (1)

Use equation (1) when making comparisons between groups for a continuous indicator collected from a non-clustered sample.

When making comparisons between groups for a **continuous indicator collected from a non-clustered sample**, use equation (1):

$$n^* = \frac{2(t_\beta + t_{\alpha/2})^2 SD^2}{\delta^2} \qquad (1)$$

where

- t_β (beta) is the critical value from the left tail of the inverse t-distribution[8] with (n^*-1) degrees of freedom.[9] Typically, the t_β critical value chosen is 80%, representing the sample's power. Thus, there is a 20% probability of not finding a difference from the intervention despite there being one (also known as a Type II error). The necessary critical value can be obtained from an appropriate t-table or the Excel T.INV command (as used in Appendix 2).
- $t_{\alpha/2}$ (alpha) is the two-tail critical value from the inverse *t*-distribution. Typically, the t_α value chosen is 5% and represents the significance level. With this value, there is a 5% chance of rejecting the null hypothesis (usually of no difference between periods/ comparison groups) when it should not be dismissed (also known as a Type I error). The necessary critical value can be obtained from an appropriate t-table or by using the Excel T.INV.2T command (as used in Appendix 2).
- SD is the indicator's standard deviation. If this is unknown, please see Appendix 1.4 for a list of possible values and a guide on calculating the standard deviation from existing data. Obtaining data from other projects within the same country program, region, or globally and conducting a desk review for standard deviations for similar indicators from previous studies may be helpful.
- δ (delta) is the targeted change in the indicator due to programming. Typically, δ is the difference between the indicator's baseline and life-of-project target values.

To work through a real-life example, at the project start-up stage, the United States Department of Agriculture (USDA)-funded Local and Regional Food Aid Procurement Project in Honduras needed to estimate the required sample size for the indicator "Mean increase in yield for project participants" growing carrots. Through the research of outside sources, program staff estimated the mean carrot yield to be 20,324 kg/ ha, with a standard deviation of 7,848 kg/ ha (Lana 2012). Program staff then examined a range of project target values to see how the resultant sample sizes varied. The project expected large yield increases due to improved techniques and thus estimated they would only need to survey 21 farmers.

$$n^* = \frac{2(0.88 + 2.26)^2 * 7{,}848^2}{(20{,}324 * 0.35)^2}$$

where $t_\beta = 0.88$ (when starting at 10 degrees of freedom); $t_{\alpha/2} = 2.26$; SD = 7,848; and $\delta = (20{,}324 * 0.35)$. However, for a more minor change, such as a 10% mean increase in yield, the project would need to survey 236 farmers. In short, larger sample sizes are needed to detect

[8] When using a sample standard deviation (and not the true standard deviation of the entire population, which is unknown), take the critical values from the *t* distribution (and not the normal or *z*-distribution).
[9] Degrees of freedom are the number of observations used in analysis. Given that *n** is still unknown, go through a few iterations, changing the estimate of *n**, until the *n** used for finding the critical value of the *t* distribution equals the *n** recommended by the equation. An example of this iterative process is shown in the spreadsheet in Appendix 2.

smaller changes with a continuous indicator. It is often helpful to put together a table, as in the spreadsheets in Appendix 2, to examine the various options. Table 1 is an example of such a table.

TABLE 1. HONDURAS CARROT YIELD EXAMPLE

CHANGE (δ)	35%	30%	25%	20%	15%	10%
TOTAL SAMPLE SIZE (n^*)	22	29	40	61	106	237

With Table 1 in hand, MEAL staff can help identify achievable targets for which statistical changes from baseline can be detected within budget constraints.

2.2.1.2 Equation (2)

Use equation (2) when making comparisons between groups for a continuous indicator collected from a clustered sample.

When making comparisons between groups for **a continuous indicator collected from a clustered sample**, use equation (2), which adds the term, $(1 + (m - 1)\rho)$ to equation (1). This term increases the sample size to offset the effects of clustering.

$$n^* = m^* k^* = \frac{2\left(t_\beta + t_{\alpha/2}\right)^2 SD^2}{\delta^2}(1 + (m - 1)\rho) \qquad (2)$$

where

- m is the number of people sampled in each cluster.
- k is the number of clusters sampled.
- ρ is the anticipated intracluster correlation (ICC) at the project's baseline. The ICC measures how much of the variability in the indicator is due to differences between clusters vs. individuals within clusters. It is further explained in Appendix 1, including how it correlates with the design effect and a list of potential ICC values.
- t_β, $t_{\alpha/2}$, SD and δ are defined as above.

For written simplicity, equation (2) is presented as so. However, for computational ease in the examples in Appendix 2, m is solved as a function of k. Thus, users will input k (clusters) and calculate m (individuals). Note that, in the clustered case, the t-distribution has 2*(k-1) degrees of freedom. It is helpful to put the various options into a table and play with different numbers of clusters to determine the best sample size given resource constraints.

In another example, the USDA-funded McGovern-Dole International Food for Education projects in Sierra Leone and Burkina Faso wanted to see how educational spending differed between households that were/ were not members of Savings and Internal Lending Communities (SILC). Using a desk review, the project assumed $\rho = 0.28$, as found in a similar Ugandan study on the household vs. community characteristics that contribute to savings (Chowa, Ansong, and R. Despard 2014). Using mean values from a SILC study in Zambia, the team noted that SILC vs. non-SILC members spent 152% more on education expenses than the 10.47 USD mean (Noggle 2017). The standard deviation of the Zambian data was high, at 24.87 USD. The project felt the budget could support detecting a more minor change than 152% in case such a large difference was not observed in the Sierra Leone/ Burkina Faso study. They surveyed 7 SILC members in each of the 27 SILCs, allowing for a 115% increase (11.95 USD) over the Zambian study's control group mean.

SILC member in Burkina Faso next to group's security box. [Sam Phelps]

$$n^* = \frac{2(0.88 + 2.26)^2 * 24.87^2}{11.95^2}(1 + (7 - 1)0.28)$$

where $t_\beta = 0.88$; $t_{\alpha/2} = 2.26$; $SD = 24.87$; $\delta = 11.95$; $k = 7$; and $\rho = 0.28$.Table 2 is an example.

TABLE 2. SIERRA LEONE / BURKINA SILC EXAMPLE

CHANGE (δ)	15.91	14.13	13.09	12.04	10.47
CLUSTERS (k)	27	27	27	27	27
INDIVIDUALS (m)	2	3	4	7	41
TOTAL SAMPLE SIZE (n^*)	54	81	107	162	1080

As Table 2 shows, if the number of clusters is held constant at 27 and SD at 24.87, the project team chooses the trade-off in change (δ) vs. individual (m) sample sizes and, by extension, budget. Twenty-seven clusters, with 2 individuals each, are the smallest overall sample size, but they limit the size of the change the team can detect.

2.2.1.3 Equation (3)

When making comparisons between groups for a **binary indicator collected from a non-clustered sample**, use equation (3):

$$n^* = \left(p_1(1-p_1) + p_0(1-p_0)\right)\frac{\left(z_\beta + z_{\alpha/2}\right)^2}{\delta^2} \qquad (3)$$

where

- p_1 is the project's target for the indicator
- p_0 is the baseline value
- z_β (beta) is the one-tailed critical value from the inverse normal distribution. Typically, the z_β critical value chosen is 80%, representing the sample's power. The necessary critical value can be obtained from an appropriate z-table or the Excel NORM.INV command (as used in Appendix 2).
- $z_{\alpha/2}$ is the two-tailed critical value from the inverse normal distribution. Typically, the z_α value chosen is 5%, representing the sample's significance level. The necessary critical value can be obtained from an appropriate z-table, or by using the Excel NORM.INV command (as used in Appendix 2).
- δ (delta) is the targeted change in the indicator and is $p_1 - p_0$.[10]

Use equation (3) when making comparisons between groups for a binary indicator collected from a non-clustered sample.

To demonstrate, the USDA-funded International McGovern-Dole Food for Education project in Burkina Faso wanted to survey mothers to measure the indicator "Percent of participants of community-level nutrition interventions who practice promoted infant and young child feeding (IYCF) behaviors." They anticipated the baseline value to be 40% and set the final target to 55%. The team concluded they would survey 171 mothers of children under 2.

$$\mathrm{n}^* = \left(0.55(1-0.55) + 0.40(1-0.40)\right)\frac{(0.84+1.96)^2}{0.15^2}$$

where p_1 = 0.55; p_0 = 0.40; $z_\beta = 0.84$; $z_{\alpha/2} = 1.96$; and δ = 0.15.

2.2.1.4 Equation (4)

When making comparisons between groups for a **binary indicator collected from a clustered sample**, use equation (4) which adds the term, $(1 + (m-1)\rho)$ to equation (3). This term increases the sample size to offset the effects of clustering.

$$n^* = m^*k^* = \left(p_1(1-p_1) + p_0(1-p_0)\right)\frac{\left(z_\beta + z_{\alpha/2}\right)^2}{\delta^2}(1 + (m-1)\rho) \qquad (4)$$

Use equation (4) when making comparisons between groups for a binary indicator collected from a clustered sample.

where all parameters are defined as above.

To use another example from the Burkina Faso USDA team, they were using a performance evaluation to determine the "Number of teachers/educators/teaching assistants who demonstrate use of new and quality teaching techniques or tools." To calculate the sample size needed at baseline, the team used final evaluation data from a previous phase to calculate the relevant ICC of

[10] Note that, in the binary case with treatment and control groups, the only time that an even, 50/50 split between treatment and control groups is optimal is when $p_0 = 1 - p_1$, for example, when the baseline value is anticipated at 40%. The anticipated final evaluation value in the treatment group (target value) is 60%. See Section 4.4 on how to determine the optimal split when $p_0 \neq 1 - p_1$.

0.44. In their IPTT, they estimated the baseline value at 49%. Their target for the indicator was 65%. The team concluded they needed to sample two teachers in 105 schools to detect a statistical change from baseline to final.

$$n^* = \left(0.65(1-0.65) + 0.49(1-0.49)\right) * \frac{(0.84+1.96)^2}{0.16^2}(1+(2-1)0.44)$$

where $p_1 = 0.65$; $p_0 = 0.49$; $z_\beta = 0.84$; $z_{\alpha/2} = 1.96$; $\delta = 0.16$, $m = 2$; and $\rho = 0.44$. Again, especially with clustered designs, it is helpful to put together a table to find the best combination of clusters and number of people per cluster to fit resource constraints, as is done in Table 3:

TABLE 3. BURKINA FASO TEACHER EXAMPLE

CHANGE (δ)	16%	16%	16%	16%	16%
CLUSTERS (k)	146	105	92	85	81
INDIVIDUALS (m)	1	2	3	4	5
TOTAL SAMPLE SIZE (n^*)	147	211	276	340	405

As Table 3 shows, if the desired change of 16% is held constant, the project team chooses the trade-off in cluster (k) vs. individual (m) sample sizes. One-hundred forty-six clusters, with 1 individual in each, is the smallest overall sample size. However, traveling to 105 schools more closely aligned with other indicators that required visits to those same schools, thus reducing overall data collection costs.

2.2.2 Single groups

If a project team is certain they only need to assess a context for one group and does not anticipate using the data to detect changes over time, then equations (5-8) should be used. This might be appropriate, for example, for a needs assessment in which no statistical comparisons between time points, geographic areas, or gender will be made.

Equations (5-8) are very similar to equations (1-4), but when the change between comparison groups is undetected, the δ term is no longer relevant. The equations do not include a β term, as there is no longer a concern about Type II errors (not finding a difference despite there being one).

Equations (5-8) follow very closely to Sullivan (2019), and are the same as those used in the Feed the Future guide for implementing partners (Stukel 2018b). See Appendix 6 for further comparison of this guide to the Feed the Future guide.

If collecting data that will used to analyze data for a single group, use equations (5-8)

- For a continuous indicator collected from a non-clustered sample, use equation (5)
- For a continuous indicator from a clustered sample, use equation (6)
- For a binary indicator from a non-clustered sample, use equation (7)
- For a binary indicator from a clustered sample, use equation (8)

2.2.2.1 Equations (5 and 6)

There are no longer two sample SDs; thus, the 2 in the numerator of equations (1-2) is no longer needed. However, the equations include a margin of error (E) to estimate the population mean within a meaningful range. For example, if the mean weight is assumed to be 60 kg and E is set to

30 kg, then the sample would estimate the adult population weight range as 30-90 kg, which is not meaningful. Instead, set E to 5 kg to estimate the true adult population weight as 55-65 kg.

Use equation (5) for single group assessments of a continuous indicator collected from a non-clustered sample. Equation (6) is the corresponding clustered version.

Equation (1) for continuous indicators thus becomes:

$$n^* = \frac{t_{\alpha/2}^2 SD^2}{E^2} \qquad (5)$$

Multiply equation (5) by $(1 + (m - 1)\rho)$ for use with clustered samples.[11] This term increases the sample size to offset the effects of clustering.

$$n^* = m^* k^* = \frac{t_{\alpha/2}^2 SD^2}{E^2}(1 + (m - 1)\rho) \qquad (6)$$

2.2.2.2 Equations (7 and 8)

In equations (3-4) for binary indicators, only the estimated probability for the indicator in the assessed population is needed, and thus there is not a p_1 and p_0 term. Note that, as described in Appendix 3, variance with a binary indicator is greatest when $p = 50\%$; this is also when the sample size for equations (7-8) is largest. Without other data, calculate the sample size for equations (7-8) with $p = 50\%$.

Use equation (7) for single group assessments of a binary indicator collected from a non-clustered sample. Equation (8) is the corresponding clustered version.

Equation (3) thus becomes:

$$n^* = \big(p(1 - p)\big)\frac{z_{\alpha/2}^2}{E^2} \qquad (7)$$

Multiply equation (7) by $(1 + (m - 1)\rho)$ for use with clustered samples, as this term increases the sample size to offset the effects of clustering.

$$n^* = \big(p(1 - p)\big)\frac{z_{\alpha/2}^2}{E^2}(1 + (m - 1)\rho) \qquad (8)$$

2.2.3 Summary.

Section 2.2 provided equations (1-4) to calculate the minimum sample size needed to detect a statistical difference between two comparison groups (e.g., baseline and final; treatment and control, etc.). The appropriate equation depended on whether the indicator was a) continuous or binary and b) collected from a clustered sample.

Suppose project teams conduct simple assessments and do not intend to detect statistical differences between comparison groups (e.g., gender, geographic area, control/ treatment, baseline/ final, etc.). In that case, it is appropriate to use equations (5-8) in lieu of equations (1-4). Depending on the margin of error or estimated probability chosen, this may result in smaller calculated sample sizes.

Before finalizing necessary sample sizes, project teams should review section 3, to determine if any additional criteria apply to their situation.

Before finalizing necessary sample sizes, project teams should review Section 3 to determine if any additional criteria apply to their situation. If so, they will need to increase their sample size further.

[11] As with equation (2) above, for written simplicity, equation (6) is presented as so. However, for computational simplicity in appendix 2, m is solved for as a function of k. In this case, the t distribution has k-1 degrees of freedom.

After finalizing calculations from sections 2 and 3, if project teams are concerned that the calculated sample size is too large, given budget constraints, review Section 4. It may be possible to reduce the sample size without affecting the size of the change to be detected.

3. Required review: Items that will increase the sample size

Equations (1-8) present the minimum size needed to detect a statistical change over time. Section 3 provides a brief overview of additional considerations that may be necessary when determining the final sample size project teams should use. These are data losses, attrition (panel datasets), specific sub-populations, and large indicator changes over time.

3.1 Data loss

CRS recommends assuming 5% data loss.

Suppose project teams anticipate some collected data will be lost due to tool, enumerator, or data entry error. In that case, they should collect data from a larger sample than the calculated minimum. CRS recommends assuming 5% data loss, but project teams should consult with colleagues in their country program on previous experience with data loss.

3.2 Attrition

If project teams anticipate some collected data will be lost due to individuals or clusters not being trackable from baseline to final (attrition)[12]; then, they should collect data from a larger sample than the calculated minimum. This issue will likely be more severe when collecting a panel dataset (where the same individual is tracked over time) vs. a repeated cross-section (usually when the same cluster is tracked over time, but individuals within the cluster vary).

Because the same individuals are surveyed at least twice for panel datasets, CRS recommends increasing the sample size by an additional 10% (above data loss concessions made above), although there are some exceptions.[13] Project teams should consult with colleagues in their country program on previous experience with attrition rates.

3.3 Indicator-specific sub-populations

Be cognizant that, for specific indicators, project teams may not be able to identify the individuals with the necessary criteria before data collection. For example, the CRS Global Result indicator "Prevalence of children 6–23 months receiving a minimum acceptable diet (MAD)" only applies to caregivers of children aged 6-23 months. Project teams may not have a detailed list of these caregivers before data collection (especially at baseline). They may thus need to over-sample the target population to sample enough caregivers to attain the identified sample size. Project teams should consult with colleagues in their country program on previous experience with indicator-specific sub-populations.

For example, if project teams know that approximately 1 out of 3 households in their participant communities have children aged 6-23 months, they could increase the sample size by 300% over

[12] Tracking beneficiaries over time (panel data) may help to reduce the overall sample size needed. See Section 4.5 for more details.

[13] The MIRA study collected panel data on the same HHs for over 24 months. Attrition was 3% - 5%; this may be due to the frequent (monthly) follow-up and CRS use of embedded enumerators that reside in the surveyed communities.

the calculated size. Enumerators may only ask the MAD-indicator-related survey questions to households once they have confirmed they have a child aged 6-23 months in their household.

3.4 Large changes from baseline to endline

When comparing groups in equations (1-4), the calculated sample size may be too small to make each data point meaningful if a large change is estimated between the groups.

In this example, it is estimated that 80% of farmers will adopt a new technique (once demonstrated to them) because it is easy to adopt and is a big improvement over current practice. Based on project design assessments, only 10% of farmers are estimated to use the technique. Due to this large change from baseline to endline and using equation (3), only five farmers must be sampled at baseline and endline to detect the 70% change over time. However, using equation (7) and adjusting the margin of error until only 5 farmers are sampled, we see that the baseline value will have a 27% margin of error. Thus, if the baseline sample mean is 10%, we can only say that the true baseline value is between 0% and 37%. For this reason, it is preferable to use equation (7) to calculate the sample size, estimating the baseline value as 10% with a 10% margin of error. This recommends sampling 35 farmers at baseline. At the endline, we'll need to survey 62 farmers to maintain the 10% margin of error. In this example, if project teams want to speak to the same farmers at baseline and endline, you should also use the larger endline sample size at baseline.

If project teams are using equations (1-4), it is recommended to cross-check individual point estimates with equations (5-8).

When teams use equations (1-4) to calculate sample sizes, it is recommended that they cross-check them with their counterparts in equations (5-8). If teams are unclear about the counterpart equation, return to the Sample Size Decision Tree. Switch to the "Single Group" branch, answer the indicator type and clustering questions, and arrive at the correct equation.

3.5 Summary

If, based on a review of Section 3, project teams determine they need to increase their sample size to account for data loss, attrition, or specific sub-populations, this can be done by increasing the number of individuals or clusters sampled. The choice depends on the likelihood of a cluster ceasing to function over the project's life (public schools are unlikely to stop, but SILC or producer associations may) vs. individuals leaving that cluster. Additional "back-up" clusters or individuals should be randomly selected in the same manner that other survey respondents are chosen.

4. Other considerations

This section presents more advanced concepts and reference tools outside this guide, about which readers may need more information. This section focuses on 1) the finite population correction factor, 2) the relationship between indicator baseline values, targets, and sample sizes, 3) stratifying samples, 4) binary indicators in impact evaluations, and 5) panel datasets and sample size calculations.

4.1 The Finite Population Correction (FPC) factor

If concerned that the calculated sample size is too large for budget constraints, see if it qualifies for reduction by the finite population correction (FPC) factor. The FPC is used when the sample size is large compared to the population size. Use the FPC if the calculated sample is greater than 5% of the population for which the indicator is being collected (regardless of it being continuous or binary). The theoretical FPC is $1 - \left(\frac{n}{N}\right)$, although it is sometimes written as $\frac{(N-n)}{N}$, where n is the calculated sample size, and N is the population size. The FPC is then multiplied by the respective confidence interval from which the sample size equations are derived (confidence intervals are further discussed in Section 6.2.6). We do not show the algebra here, but equation (9) is used to adjust the initial sample size by the FPC for both continuous and binary indicators; note it follows Thompson (2012):

$$n_{\text{FPC}} = \frac{1}{\frac{1}{n^*} + \frac{1}{N}} \qquad (9)$$

with *n** being the initial sample size calculated and *N* defined above.

The FPC should be applied to the sample size calculated before adjusting for data loss, attrition, or specific sub-populations described in Section 3.

It is not recommended to apply the FPC if assuming external validity, such as with impact evaluation or research activities.

Note that, in the economics literature, the FPC is usually ignored because researchers assume that the sample size is small relative to the entire population (Cameron and Trivedi 2005)This would be especially true when assuming external validity (generalizability beyond a specific project) with impact evaluation or research activities. This usually occurs with an experimental design that compares treatment group(s) to a control group. In such cases, applying the FPC is not recommended.[14]

4.2 Set achievable targets and be cognizant of their associated data collection costs

Understanding the implications for MEAL budgets during the project design stage is essential. Data for donor standard indicators must be collected and reported, and adequate resources should be allocated to detect changes in these indicators over time.

However, custom indicators are somewhat at the discretion of project teams. For example, during a recent education project design in Togo, the team wanted to collect school-age children's body mass indices (BMI). It is difficult to find this data for any country, and the project design team guessed it could see a small (3-5%) change in this indicator over the 5-year project but was unsure.

[14] If reporting confidence intervals around a sample mean, and if the FPC criteria apply, it should also be used to adjust the confidence interval. See Section 6.2.7.

Given the small change for a continuous indicator, the required sample size was 50% larger than any other project indicator.

The project design team thus decided to collect the data through a special study while MEAL teams were collecting other data from students, using the smaller sample size recommended for other indicators. Although detecting a statistical change over the project's life might not be possible, the data should prove helpful in informing future projects in Togo and potentially other CRS education projects globally. If a larger change than anticipated is achieved unexpectedly, the project team can detect a statistical difference from baseline to final. Thus, the project team removed this indicator from the project's IPTT, which had the added benefit of not committing the project team to a targeted change that would be very expensive to detect if they could do so at all.

4.3 Stratification

Stratification means arranging or classifying data into groups. This has been mentioned above when discussing treatment and control groups, gender, or geographic separation. When calculating sample sizes, it is vital to think about any stratification project teams will do with the data for two reasons:

If project teams would like to make statistical comparisons between strata, then the sample size needs to be increased to account for this.

1) **Statistical Comparisons Between Strata.** If project teams would like to make statistical comparisons between strata, then the sample size needs to be increased to account for this. As described in Section 1, this is typically done by multiplying the final sample size by the number of strata to be analyzed. For example, if testing boys' vs. girls' educational outcomes (2 strata), and the sample size is 5 children in 50 schools, speak to 5 boys and 5 girls in each school.

2) **No Statistical Comparisons Needed.** If project teams need to stratify the data but do not want to make statistical comparisons between the strata (this is often done for gendered disaggregation of data), then the sample size need not increase. For example, suppose the team would like to know the yields of both male and female producers. In that case, they can randomly select male producers from a list and choose separately female producers to ensure the representation of both genders in the sample without increasing the overall sample size.

Avoid the need for sample weights when stratifying a sample by using fractional interval systematic sampling.

Sample weights, described in Section 6.2.1, must account for any sample stratification. This is true even when no statistical comparison is made between strata.

- One potential way to address the need for weights is to use fractional interval systematic sampling. More details on this sampling strategy can be found in Section 9.4.2 of Stukel (2018b).

When making statistical comparisons between strata and clustering the sample, increasing the number of clusters or individuals within each cluster is possible. Maintain a consistent sample size within each cluster and ensure sample sizes are accounted for in stratification. To build off the example in point 1) above, if testing differences between geographic regions (2 strata), speak to 5 children in each of 50 schools in Region A and do the same in Region B.

If testing differences by geographic region and gender (4 strata), speak to 5 boys and 5 girls in 50 schools in Region A, and do the same in Region B. In this case, ensure the expected differences from baseline to endline for the overall sample with 10 children per cluster is satisfactory, as well as the expected differences only for girls and only for boys. Use the larger calculated sample size of the various options.

4.4 Use unequal treatment and control group sizes with binary indicators

When using an impact evaluation to measure a binary indicator[15] from a clustered sample, allowing for unequal numbers of clusters in the control vs. treatment group is more efficient. With binary indicators, the only time that a 50/50 split between treatment and control groups is optimal is when $p_0 = 1 - p_1$. For example, the baseline value is anticipated at 40%, and the treatment group's final value is anticipated at 60%.

Although the split can be left at 50/50, if conducting an impact evaluation (or research) with comparisons between treatment and control groups, it may be more cost-efficient to have an uneven split. This may be especially true if there is a considerable cost to the treatment itself; the treatment group can be the smaller of the two groups.

Given the rarity of this occurrence in the CRS context, this guide recommends that project teams refer to equations 17, 18, and 21 (and their accompanying Excel spreadsheet) in McConnell and Vera-Hernández (2015) to calculate the optimal split between treatment and control groups.

4.5 Use panel datasets

If tracking the same individual over time, project teams can gain statistical power by using baseline values when analyzing final evaluation data. Calculating sample sizes that explicitly account for the additional statistical power provided by panel datasets requires project teams to have additional information for the sample size equations, which may not be available. The additional information is:

To calculate sample sizes that account for the additional statistical power provided by panel datasets, see McConnell and Vera-Hernández (2015).

- For continuous indicators, the ICC at both the individual and cluster level. If interested in this approach, refer to equation (16) in McConnell and Vera-Hernández (2015).
- For binary indicators, the baseline value of the outcome variable, or any other covariate, may reduce the calculated sample size. The user will need to empirically estimate how the covariate affects the indicator before calculating the sample size, and the sample size equations themselves can be computationally overwhelming for those unfamiliar with linear algebra. If interested in this approach, refer to equations (24-25) in McConnell and Vera-Hernández (2015). Appendix 3 of CRS Samples provides an example R code for equations (24-25).

In lieu of using sample size equations specific to panel datasets, use equations (1-4) of this guide, as appropriate. They do not account for the additional statistical power provided by panel datasets but will ensure the sample size is sufficiently large.

4.6 Summary

This section included special topics about which advanced users, likely technical advisors, should be aware regarding sample size calculations. It included an overall discussion about the finite population correction factor, the relationship between sample sizes and project targets, sample stratification, binary indicators in impact evaluations, and calculating sample sizes for panel datasets.

[15] Note that, with continuous indicators, it is less efficient to have an uneven split between treatment and control groups. See equation 3 in McConnell and Vera-Hernández (2015).

5. Sample size myths

Within the international development field, there are two widely held misunderstandings regarding sample size calculations: 1) that sample sizes depend on the underlying population size and 2) that binary indicators require larger samples than continuous ones. Both myths are debunked in this section.

5.1 Sample sizes depend on the underlying population size

After reviewing this guide, note that the only time the underlying population size is considered is when the FPC is deemed appropriate (Section 4.1). The factors influencing sample size calculations are 1) the size of the change to be detected (or acceptable margin of error) and 2) the number of comparison groups. Only when the sample size is greater than 5% of the underlying population should we consider reducing it by the FPC factor.

5.2 Binary indicators require larger sample sizes

Some data can be "reasonably" converted from binary to continuous, and vice-versa. There is an oft-stated myth that using the continuous version is preferable, as it will require a smaller sample size (Frost 2020). In comparing the basic equations for continuous and binary variables, equations (1) and (3), respectively, each has different parameters, and there is no mathematical proof that one would always result in a smaller or larger sample size.

$$n^* = \frac{2\left(t_\beta + t_{\alpha/2}\right)^2 SD^2}{\delta^2} \quad (1) \qquad n^* = \left(p_1(1-p_1) + p_0(1-p_0)\right)\frac{\left(z_\beta + z_{\alpha/2}\right)^2}{\delta^2} \quad (3)$$

To provide an example using test scores from primary school students in Sierra Leone, the project's donor-required standard indicator is the percentage of children passing the reading test, which is a binary indicator. The project team anticipated the baseline value to be 41% and set a target of 58% and thus wanted to measure a change of 17%.

Students answering 3 of 5 questions correctly are considered to have passed the exam. The anticipated baseline value and target could theoretically be converted to a score. If 41% of students scored at least 3 on the test (and all others scored zero), the average baseline score would be 3*0.41 = 1.23 per student; the target would thus be 3*0.58 = 1.74 per student, and the change would be 0.51. Using the actual test score data from this example, the standard deviation is 1.83. In this example, equation (1) for the continuous version of the indicator calculates the larger sample size.

$$n^* = \frac{2(0.88 + 2.26)^2 1.83^2}{0.51^2} = 205$$

$$n^* = \left(0.58(1-0.58) + 0.41(1-0.41)\right)\frac{(0.84 + 1.96)^2}{0.17^2} = 132$$

It is important to remember that continuous and binary indicators have different probability distributions and, thus, variances. For this reason, they require different equations, and blanket statements about the resultant calculations cannot be made. The parameters needed for each equation are different because they measure fundamentally different things.

5.3 Summary

Always use the appropriate equation for the indicator type, as the final statistical tests performed at the analysis stage will depend on it. Ultimately, the goal with sample size calculations is to have a sufficiently large sample to detect statistical changes at the analysis stage, and the underlying population size is not a factor in these calculations.

6. Sample design and analysis

Although this guide focuses on sample size calculations, it is important to understand how the chosen method of sample selection, made prior to determining which sample size equation to use, impacts analysis. This section provides a brief overview of random sampling and sample selection bias, population proportional to size sample selection, and the use of sample weights at the analysis stage.

6.1 Sample Selection

This section describes the importance of using random samples and the challenges of using the Probability Proportional to Size (PPS) methodology.

6.1.1 Use random samples and document any sample bias due to non-random sampling

Representative samples should always be selected randomly from a pre-populated list or a rapid census, and the probability of selecting a person for the sample should be known.[16] CRS has internal references for various ways of selecting a random quantitative sample. (Culligan et al. 2019) Random sample selection is critical to achieving external validity; it will be challenging to publish evaluations or research findings from a non-random sample externally.

If missing 5% or more responses to an individual question, in the limitations section of the study report, describe any sources of bias due to these omissions.

However, often due to resource constraints, non-random sampling and, thus, selection bias does occur. This may be due to security constraints that prevent study teams from reaching an off-limits area or when the rosters from which individuals or clusters are randomly selected are outdated. It would prove too costly or impossible to locate those randomly selected. If missing 5% or more respondents or responses to an individual question (Cameron and Trivedi 2005), in the limitations section of the evaluation report, describe any sources of bias due to these omissions as well as possible.

For example, if students are not present in school on the day they are meant to be surveyed, how do absent students differ from those present? Does a t-test of means show that the proportion of key groups (gender, ethnicity, geographic area)[17] in the sample is the same as those that were not included? If not, how might the sample be biased? How else might students not present that day be different? Might they not perform as well on literacy tests, etc., because they might frequently miss school?

Another example is if yield can be measured for some farmers but not others because they a) did not plant the crop being surveyed or b) planted the crop but did not harvest it?

Imagine other scenarios in which this could occur and be thoughtful about what can be said about the differences between those that could/ could not be surveyed.

[16] For this reason, the use of a "random walk" for household selection is not recommended, unless the total number of households in a community is known.

[17] The analyst may not have much information about students not present. However, based on student names and school locations, they might at least have this information.

6.1.2 Population Proportional to Size (PPS) cluster selection may not be appropriate in the CRS context

PPS is one method for selecting study clusters. It is commonly used to account for the size of clusters when selecting them in the first stage of data collection, in which every individual in every cluster has an equal probability of being selected into the sample. If, in the second stage, a simple or systematic random sample is used to select survey respondents, then the sample is "self-weighting," and no sample weights need be applied at the analysis stage.

PPS cluster selection is only self-weighting if simple or systematic random sample selection is used in the second stage.

Analysts of data collected via a PPS-selected sample should understand four things:

1) If the sample was stratified (as described in Section 4.3), or if a simple or systematic random sample was not used in the second stage, then the sample is not self-weighting, and sample weights must be used (see Section 6.2.1 for a discussion of sample weights).

2) the size measure must be the same as the sampling unit used at the analysis stage to use PPS. Note that different units (households, producers, caregivers, etc.) are often needed for other indicators in the same project. Thus, each sampling unit will need to draw a different sample. For example, using the total number of households in a village as the "size" in PPS and then using caregivers of children aged 6-23 months as the sampling unit is incorrect. If PPS is used in this example, it is necessary first to know how many caregivers of children aged 6-23 months are in each village and use that as the "size."

3) With PPS, the Hansen-Hurwitz or Horvitz-Thompson estimators should be used to estimate the sample mean. In addition, those estimators should be used when calculating the variance in any regression models (Hansen and Hurwitz 1942; Horvitz and Thompson 1952). This point is not typically addressed in other sampling guides.

4) Also, the size measure should be accurate when using PPS. Otherwise, it will over- or underestimate the sample variance, as compared to a simple random selection of clusters (Thomsen, Tesfu, and Binder 1986). Even if baseline size measures are accurate, if a repeated cross-section is used in the same clusters at final evaluation and the "size" of the clusters changes notably over time, the same issue of misestimating the sample variance will occur.

Feed the Future (FtF) recommends using PPS in its sampling guide for implementers. When surveying individuals through producer groups, it recommends surveying every producer in the selected group; this would avoid the issues described in points 1) and 2) above. Regarding point 2), household surveys should be used as a size measure, and an updated list of all individual project participants is recommended. It then uses the largest calculated sample size across all project indicators to survey selected participants while acknowledging the risk of not sampling an adequate number of respondents per indicator (when the indicator only applies to a specific sub-population). (Stukel 2018a, 2018b).

BHA's emergency activity M&E guide provides three examples of using PPS to select clusters with multiple indicators. The household is always the sampling unit and a simple or systematic random sample is used in the second stage. The examples demonstrate the complexity of using PPS as a sampling methodology and should be reviewed by any CRS staff interested in using PPS for multiple indicator data collection. (*Technical Guidance for Monitoring, Evaluation, and Reporting for Emergency Activities* 2022)

Given the complexity of the analysis, concerns about measures of cluster sizes, and increased costs due to larger sample sizes for indicators, CRS staff should use PPS carefully. Instead of PPS, clusters and individuals can be selected via other forms of random sampling, and sample weights can be used in the analysis. Note, however, that if deciding where to use PPS and analysts will run

regressions on the collected data, sample weights may reduce the precision of coefficient estimates (Lee and Solon 2011). A more detailed discussion of sample weights used with regression analyses is in Section 6.2 below.

6.1.3 Summary

It is important to select random samples. Whenever this is not possible, be sure to document any sample selection bias as well as possible. If using Population Proportional to Size (PPS) cluster selection, understand its complexity and limitations before its use. If sample weights are needed (with or without using PPS), reference Section 6.2 on sample weights.

6.2 Using Sample Sizes in Data Analysis

This section describes why, when, and how to use sample weights. It also provides an overview of non-response weights and sample weights in regression analyses. It closes with a section on calculating confidence intervals and the applicability of the finite population correction factor at the analysis stage.

Sample weights are only applicable for some clustered or stratified samples.

6.2.1 Sample weights - calculation

Sample weights are only applicable when respondents have an unequal probability of being selected for a sample. This happens with some clustered or stratified samples (as with simple or systematic random sampling, everyone has the same sample selection probability). They apply to stratified samples that were increased in size to allow for analysis between strata or to samples that were stratified for organizational purposes (see Section 6.2.1).

Sample weights reflect the number of people in the population one individual represents. For example, if schools are randomly selected in the first stage of a clustered design, and in the second stage, 10 girls are randomly selected from the school's only 2nd-grade class[18] There are 30 girls in the class; each girl sampled represents 3 other girls.

Statistically, sample weights are the inverse probability of being selected into the sample and are defined as $w_i = 1/\pi_i$, where π_i is the probability of individual *i* being selected into the sample. In the above example, π_i for each girl is 10/30; thus, her w_i is 3.

One additional point to ease calculations: In this example, if 50 schools were randomly selected from 200 in the first stage, then each school represents 4 other schools. Since every school has the same sample weight, the school weights can be ignored in the calculations.

If the 50 schools were stratified between two geographic regions, resulting in different school weights, then their weights must be included. For example, Region A has 110 schools, and Region B has 90. Twenty-five schools were randomly selected from Region A and twenty-five from Region B. Thus, Region A schools have a weight of 110/25, and Region B schools have a weight of 90/25.

6.2.2 Sample weights - use

Sample weights should always be used when providing univariate descriptive statistics for individual indicators, such as means/ proportions, totals, medians, etc.

However, results from regression analyses would ideally report unweighted and weighted results, and where there are differences, include a discussion of the underlying reasons. For example, observations from a school that has 90 second graders vs. 30 will carry 3 times the weight; if there are heterogenous project effects for large vs. small schools (e.g. larger schools have a higher

[18] This example references a USAID Education indicator that specifies 2nd grade students. Thus, all non-2nd grade classrooms would be excluded from the survey, and do not need to factor into sample weights.

teacher/ student ratio, this lack of student attention may result in poorer educational outcomes, etc.), then the conditional means might be different for weighted vs. unweighted analyses (Solon, Haider, and Wooldridge 2015). Sample weights may also reduce the precision of coefficient estimates (Lee and Solon 2011).

6.2.3 Weighted means/ proportions and totals

The equation for calculating a univariate weighted mean is:

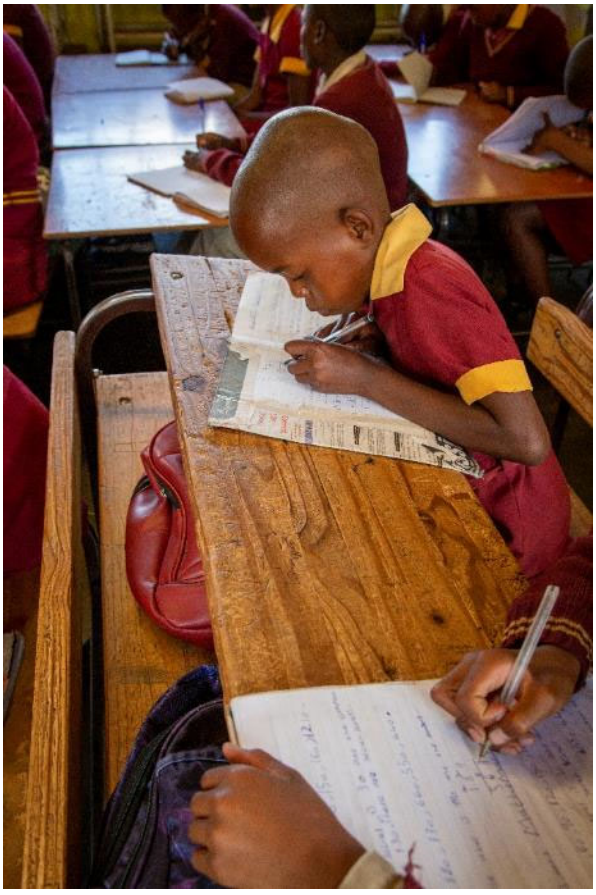
$$\bar{y} = \sum_{i}^{I} y_i * w_i \Big/ \sum_{i}^{I} w_i \qquad (10)$$

where *y* is our indicator of interest for individual *i* with weight *w*.

As a simple example, use a binary indicator (pass = 1/ fail = 0) for 5 students each from 2 classrooms (Table 4). Since each classroom has a different number of students, those in classroom A (30 students) have a weight of 30/5 = 6, and those in classroom B (40 students) have a weight of 40/5 =8.

TABLE 4. SAMPLE WEIGHT EXAMPLE DATA

Student ID	y_i	w_i	$y_i * w_i$
1	1	6	6
2	1	6	6
3	1	6	6
4	1	6	6
5	0	6	0
6	1	8	8
7	0	8	0
8	1	8	8
9	0	8	0
10	0	8	0
SUM	**6**	**70**	**40**



Students in primary school in Lesotho. [Dooshima Tsee]

Thus, the weighted proportion of students who passed (using equation (10)) is:

$$\hat{y}_w = 40/70 = 57\%$$

whereas the simple proportion is $\hat{y} = \sum_i^I y_i / n$ = 6/10 = 60%[19].

When using a representative sample to extrapolate to a larger population, multiply $\hat{y}_w$ by the total population. In this example, 0.57*(30+40) = 40 students are estimated to have passed the exam compared to the 0.60*(30+40) = 42 students that would have been estimated using the unweighted sample.

6.2.4 Non-response weights

Non-response weights are used when some of the intended sample did not respond to the survey. This non-response could be because participants refused to participate, could not be located, or the data was unmeasurable (e.g., yield from a crop that was never harvested).

Samples are thus weighted based on observable characteristics of respondents and non-respondents (such as gender, age, geographic location, etc.), in addition to their probability of selection as outlined in the previous section. For example, female respondents would be weighted to represent female non-respondents, or older respondents would represent older non-respondents. This guide does not provide details on how to calculate and use non-response weights, but Raab (2009) and National Research Council (2002) explain the practical application in detail.

Use non-response weights with caution. Do not assume that survey respondents can adequately replace non-respondents.

Please note that non-response weights should be used with caution. Any survey non-response that is not random creates a biased sample, and that bias should be documented (see Section 6.1.1). Analysts cannot assume that actual survey respondents can adequately replace non-respondents.

To continue the example from Section 6.1.1, why can yield only be measured from some farmers? Is the primary reason that some farmers did not harvest, and thus, there was no yield to measure? Did they not harvest because they could not irrigate in a drought year or use drought-resistant agricultural inputs? In this case, using non-response weights to give greater weight to respondent farmers who harvested (assuming they can represent farmers who did not harvest) would exacerbate the sample bias.

6.2.5 Clustered or stratified samples and regression analysis

When reporting weighted conditional means from regression analyses, weighted values should use the appropriate weighted counterpart (e.g., weighted least squares, weighted maximum likelihood, etc.).

Additionally, because observations within a cluster are likely correlated, coefficient standard errors should always be clustered (Cameron and Miller 2015). Statistical packages have functions; the appropriate function will vary depending on the analysis method.

Control for any sample stratification or randomization in regression analyses using binary variables for each stratum or unit of randomization (excluding one to avoid the dummy variable trap).

6.2.6 Confidence Intervals

The right-hand side of the equation (11) is also known as the margin of error.

For reference, below are the equations for calculating the confidence intervals (CI) around a sample mean. For any random sample, the following equation applies:

$$\text{Confidence Interval} = \text{mean} \pm SE * t_{\alpha/2} \qquad (11)$$

SE is the sample mean's standard error and is defined in the next paragraph. For binary indicators, $z_{\alpha/2}$ should be used instead of $t_{\alpha/2}$, and both are described in Section 2.2.1. Typically, when

[19] In statistical notation, the ^ above y is used for proportions, whereas ¯ above y is used for means.

reporting confidence intervals, α = 0.05 in social sciences. The right-hand side of equation (11) is also known as the margin of error.

Note that the standard errors of the sample mean or proportion will differ for continuous (equation (12)) and binary (equation (13)) indicators.

$$SE_{Continuous} = \frac{SD}{\sqrt{n}} \qquad (12)$$

See Appendix 1 if the standard deviation (SD) equation is needed. In equations (12) and (13), *n* is the final sample size after data collection.

$$SE_{Binary} = \sqrt{\frac{p(1-p)}{n}} \qquad (13)$$

With clustered samples, the SE should be multiplied by $\sqrt{\frac{1+\rho}{1-\rho}}$ where ρ is the ICC (Bence 1995). This should be done before inserting the SE into equation (11). This is true for both continuous and binary indicators.

6.2.7 FPC

If the Finite Population Correction (FPC) factor applies to the sample (see Section 4.1), the SE should also be multiplied by it. As a reminder, the FPC is $1 - \left(\frac{n}{N}\right)$.

6.2.8 Summary

Section 6.2 explained that sample weights essentially allow observations from a single cluster to represent others not surveyed in the same cluster; observations from larger clusters will thus be given greater weight. Sample weights should be used in summary statistics, but regressions should report results from both weighted and unweighted samples.

Section 6.2 also provided an overview of non-response weights and how to use weighted samples in regression analyses; both sub-sections include links to other references for more information. It closed with the equations needed to calculate confidence intervals for binary and continuous indicators and noted the necessary modification for indicators collected from clustered samples. If used in the sample size calculations, the finite population correction factor should also be applied to confidence intervals at the analysis stage.

Appendix 1. Intracluster correlation coefficient

The Intracluster Correlation Coefficient (ICC) only applies to clustered designs and indicates how much of the variability in the data is due to differences between clusters vs. individuals within clusters. To get an accurate picture of the population, if individuals within clusters are like each other, it is best to survey fewer individuals within each cluster and a larger number of clusters. Otherwise, the similarity of responses within a cluster will magnify the differences in responses of individuals between clusters, leading to greater standard deviations (Killip, Mahfoud, and Pearce 2004).

A1.1 Design effect vs. ICC

The design effect is dependent on the ICC and the number of individuals per cluster.

Whenever a design effect is reported for a study or when a design effect is used in an equation, the authors should clarify the underlying equation used. In most cases, when working with sample size equations, the underlying equation is:

$$\text{Design effect} = 1 + ((m - 1)\rho) \qquad (14)$$

where ρ and m are the ICC and number of individuals per cluster, respectively. Note that the design effect depends on both variables; thus, equations that use a design effect without specifying m exclude critical information.

The ICC, however, is not dependent on the number of individuals surveyed, thus making it more portable across survey designs (Stukel 2018a). See below for more information on calculating an ICC from survey data. Appendix 1.4 lists already calculated ICC values for selected standard indicators for major donors to CRS projects.

This guide encourages self-calculation of ICCs from recent data for similar indicators in the same operating area.

Typically, baseline values of the ICC are used, but the ICC is likely to increase over time. This is because project activities are often at the cluster level (e.g., teacher training is done for all teachers in each school, all farmers in a producer group, etc.). Thus, individual results will likely become more correlated within clusters after receiving project interventions. Given that CRS has been operating in most of its country programs for many years, it is likely that one project is a follow-on to another or a follow-on to another NGO, potentially in the same clusters. For this reason, this guide encourages self-calculation of ICCs from recent data for similar indicators in the same operating area. In addition, Handa et al. (2018) have shown the differences in ICCs for the same indicator across geographic regions, further emphasizing the importance of using project—or country-specific ICCs.

A1.2 Calculating the ICC

This section shows how to calculate an unconditional sample ICC (with no covariates).

The sample ICC is the ratio of between and within cluster variance:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \qquad (15)$$

An ICC value of 1 would mean that differences between clusters could explain all the variation in the data. Thus, surveyors would need to visit many clusters but only one individual per cluster. An

ICC near 0 would mean that differences between individuals could explain all the data variation. Thus, surveyors would visit fewer clusters but survey all the individuals in each.

A1.2.1 General variance.

As mentioned above, the equations for calculating ICCs differ between continuous and binary indicators. Note that variance for a discrete random (continuous) indicator is defined as

$$\sigma_{\text{Continuous}} = \frac{\sum(y - \bar{y})^2}{n} \qquad (16)$$

where y is an individual observation, $\bar{y}$ is the mean of all observations, and n is the number of all observations. Variance for a binary indicator is:

$$\sigma_{\text{Binary}} = p(1 - p) \qquad (17)$$

Where p is the proportion of the population for which the condition is true. Note that variance for a binary indicator is greatest when the indicator is true for exactly 50% of the population. To see this, plug in 0.35, 0.50, and 0.70 for p. Note the respective variances are 0.23, 0.25, and 0.21.

A1.2.2 ICC for continuous indicators

For **continuous data**, between cluster variance is calculated as

$$\sigma_b^2 = \frac{\sum_c^C n_c(\bar{y}_c - \bar{y})^2}{C - 1} \qquad (18)$$

where $\bar{y}_c$is the mean of the individual-level data from the c[th] cluster. $\bar{y}$ is the mean of all observations (DataCamp 2019).[20] *C* is the total number of clusters. Thus, the numerator adds the variance for each cluster, weighting it by the number of observations per cluster. The denominator divides by the number of clusters, thus providing the mean between cluster variance.

Similarly, within cluster variance is calculated as

$$\sigma_w^2 = \frac{\sum_{i1}^{I1}(y_{i1} - \bar{y}_1)^2 + \cdots + \sum_{ic}^{IC}(y_{ic} - \bar{y}_c)^2}{n - C} \qquad (19)$$

Where y_{i1} denotes the observation for individual *i* in cluster 1, and y_{ic} denotes the observation for individual *i* in the c^{th} cluster. Thus, the numerator adds the sum of the variances within each cluster. The denominator divides by the number of observations (reduced by the number of clusters), thus providing the mean within cluster variance.

Note that in appendix 2, the "ICC_example" tab contains a worked example for a continuous indicator (second-grade literacy scores - Koinadugu, Sierra Leone), where ρ= 0.92. This same example can be found in Appendix 3 if project teams prefer to use the R statistical software.

For calculating the ICC of binary indicators, use a more robust statistical software than Excel.

A1.2.3 ICC for binary indicators

For binary indicators which follow the Bernoulli distribution, a variety of methods have been proposed (Schochet 2013; Goldstein, Browne, and Rasbash 2002). For binary indicators, it is recommended to use a more robust statistical software than Excel. The R statistical software has a

[20] For between cluster variance, use the weighted observations. If observations within a cluster have different weights (e.g., the sample was stratified by gender), use the weighted observations. Note that, if all observations within a cluster have the same weight, using the sample weights has no effect on the within cluster variance.

built-in function, ICCbin()[21] in the "aod" package that calculates ICCs for the Methods A-C as described in Goldstein. Method A uses the logistic distribution, often used when analyzing binary indicators, and is thus recommended. Appendix 3 presents example R code, using the same literacy scores as for the continuous example, but converting them to a binary (pass/ fail) indicator. In the binary version, ρ= 0.60.

A1.3 Calculating the standard deviation

For reference, the equation to calculate a sample's standard deviation (which is only used with continuous indicators) is:

$$SD = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} \qquad (20)$$

The "STDEV.S" function in Excel can also be used.

If calculating the standard deviation from a clustered or stratified sample, use the equation for a weighted standard deviation (The Statistical Engineering Division 1996):

$$SD_{\text{weighted}} = \sqrt{\frac{\sum_i^I w_i(y_i - \bar{y}_w)^2}{(n'-1)\frac{\sum_i^I w_i}{n'}}} \qquad (21)$$

where n' is the number of non-zero weights. In practice, all weights will be >0, so $n' = n$.

Sections 6.5.2.2 and 6.5.2.3 in Higgins, Li, and Deeks (2019) guide calculating standard deviations from journal articles that do not report standard deviations but do report means, confidence intervals, standard errors, and p-values.

Appendix 1.4 provides ICCs and standard deviations for most USG standard indicators.

A1.4 ICC and standard deviation values for select indicators

This appendix section provides ICC values and standard deviations for select commonly used or donor-specific indicators whose underlying data is typically collected via a representative sample. It is kept separate from the leading guide to allow for frequent updates. Please consult the cover page of Appendix 1.4 for the date of its most recent update. Readers wishing to contribute ICC values to the tables in Appendix 1.4 or would like support in calculating them should contact the author.

[21] Be sure to use the ICCbin() function from the "aod" package, and not the "ICCbin" package, as the latter does not seem to work.

Appendix 2. Sample size calculator – Excel

This appendix is an Excel spreadsheet kept separate from the leading guide. It contains ready-to-use tables for this guide's equations (1-8) and an example of calculating a continuous ICC from sample data.

Appendix 3. Sample size calculator – R

This appendix is R code, kept separate from the leading guide. It contains ready-to-use code for calculating the ICC of a binary indicator and a binary indicator with one covariate.

Appendix 4. Formal descriptions of sample size calculations

When providing sample size estimates in any document, the reader should have enough information to recreate the necessary calculations. Thus, note if the project uses a clustered design, the sample size will be calculated for each sampling frame that will be surveyed (i.e., individual students and schools, teachers, producers, etc.). Also, note the number of clusters and the number of individuals per cluster. Also, clarify the values of all parameters used (α, β, δ, ρ, p, SD, etc.) It may be easier to present some of this information in a table. See the below example from a recent proposal.

> "A two-stage cluster sampling approach will be used to select all respondents for the quantitative surveys (Table A4.1). In the first stage, schools will be randomly chosen as clusters. Then, students, teachers, cooks, caregivers, and mothers (within respective communities that feed into schools) will be selected in the second stage. The principal of each school will be interviewed as well. The equations used to determine the sample size generate the minimum sample size needed to detect a statistical difference in key outcome indicators over time. All samples will be increased by at least 5% in case of data collection or transcription errors.
>
> Sample sizes were calculated using equations (6), (19), and (22) for clustered continuous, non-clustered binary, and clustered binary outcomes, respectively, in McConnell and Vera-Hernandez (2015), using the standard 80% power and 5% significance level. Indicator-specific details are noted in individual footnotes. Some indicators were converted to percentages so that changes in mean differences could be detected over the project's life."

TABLE 5. SAMPLE SIZE PRESENTATION EXAMPLE

INDICATOR	INDIVIDUAL RESPONDENT	BASELINE (ESTIMATED)	TARGET (LIFE OF PROJECT)	INTRA-CLUSTER CORRELATION (ICC)	CLUSTERS	INDIVIDUAL PER CLUSTER
Average student attendance rate[22]	Classrooms	93%	97%	0.74	1,500	5
Percent of students demonstrating they can read	Students	21%	41%	0.43 (Duflo, Glennerster, and Kremer 2007)	44	5
Percent of caregivers who report spending time on literacy activities	Caregivers	42%	62%	0.20[23]	34	5
Percent of children 6–23 months receiving a minimum acceptable diet	Mothers of children 6-23 months	67%	79%	0.08 (Moss et al. 2018)	56	5

[22] This is a new indicator for USDA, and it is difficult to find published ICC values for it. However, based on ICC calculations from official attendance data from Sierra Leone's CRS-implemented McGovern-Dole project (Phase 4), the expected variability is based largely on schools, and not classrooms within schools, hence the large ICC. The relevant standard deviation was 0.44. Given the very large sample needed, STARS will simply use a census of all classrooms in all schools for this indicator for the baseline study and revisit these calculations using baseline data after collection.

[23] Given the custom nature of this indicator, it is difficult to find a published ICC value for it. Given that this will be household practice, an ICC value of 0.20 is assumed, which is between the school and mother-level identified ICCs above.

Appendix 5. Quick reference guide

It is best to use this appendix after familiarizing oneself the main guide.

This quick reference guide is meant to be a rapid reminder of the basic steps in the sample size calculation process. It is best used after reviewing the complete guide and does not cover as many possible scenarios.

Step 1: Compile a list of indicators to be measured (Table 6 provides an example). For each indicator, note if it is continuous or binary, the respondent, and the respondent's cluster (if data collection will be clustered). *Section 1 of the guide.*

Step 2: Add to the list, for each indicator, the change to be measured. This is often the Life-of-Project target less the expected baseline value. Still, differences between control and treatment groups may also be expected at the end of a study or differences between groups at the end of a project.

If no comparisons will be made, note the margin of error within which the indicator will be measured. *Section 2.2.2 of guide.*

Step 3: Add to the list the intra-cluster correlation coefficient (ICC) for each indicator whose data collection will be clustered. *Appendix 1 of guide.*

Step 4: Add the relevant standard deviation for each continuous indicator to the list. *Appendix 1.3 of guide.*

Step 5: Use Figure 1 to determine which equation to use for each indicator. Add this to the list. Note that if in Step 2 it was determined that no change over time will be measured, use equations (5-8) instead of equations (1-4).

Step 6: Using the above information and the Excel spreadsheet in Appendix 2, calculate the minimum number of clusters (if applicable) and respondents per cluster. Add this to the list.

Step 7: If respondents (sample frames) overlap indicators, use the largest recommended sample size per respondent type.

Step 8: Finalize the calculations by increasing the sample size by 5-20% to account for data collection errors, attrition, etc. *Section 3 of guide.*

TABLE 6. EXAMPLE PREPARATORY LIST FOR CALCULATIONS

INDICATOR	CONTINUOUS OR BINARY	RESPONDENT TYPE	CLUSTER TYPE	CHANGE TO DETECT	ICC	STANDARD DEVIATION	EQUATION	CLUSTERS NEEDED	RESPONDENTS NEEDED PER CLUSTER	FINAL CLUSTER (RESPONDENT) SIZE
Average student attendance rate	Continuous	Classrooms	School	4%	0.74	0.44	2	1,500	5	Census[24]
Percent of students demonstrating they can read grade level text	Binary	Students	School	20%	0.43	N/A	4	44	5	45 (6)
Percent of individuals demonstrating use of new safe food preparation practices	Binary	Cooks	School	35%	0.90	N/A	4	14	2	16 (2)
Percent of caregivers who report spending time on literacy activities with their school-age children in the previous week	Binary	Caregivers	School	20%	0.20	N/A	4	34	5	35 (3)
Percent of children 6–23 months receiving a minimum acceptable diet	Binary	Mothers of children 6-23 months	School	12%	0.08	N/A	4	56	5	56 (6)

[24] Note the project served fewer than 1,500 schools, thus they will need to do a census of all classrooms in all schools.

Confidential - Appendix 6. Other sample size guides

Appendix 6 should not be shared outside of CRS. This includes CRS partner organizations.

This appendix focuses on other non-CRS sample size guides. It is maintained as a separate appendix from the leading guide, containing some proprietary information to CRS.

Bibliography

Bence, James R. 1995. "Analysis of Short Time Series: Correcting for Autocorrelation." *Ecology* 76 (2): 628-639. https://doi.org/0.2307/1941218.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317-372. https://doi.org/10.3368/jhr.50.2.317.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge University Press.

Catholic Relief Services. 2023. "Monitoring, Evaluation, Accountability and Learning (MEAL) Policies & Procedures." https://www.crs.org/our-work-overseas/research-publications/monitoring-evaluation-accountability-and-learning-policies-procedures.

Chowa, Gina, David Ansong, and Mathieu R. Despard. 2014. "Financial Capabilities: Multilevel Modeling of the Impact of Internal and External Capabilities of Rural Households." *Social Work Research* 38 (1): 19-35. https://doi.org/10.1093/swr/svu002.

Culligan, Mike, Leslie Sherriff, Clara Hagens, Guy Sharrock, and Roger Steele. 2019. *A Guide to the MEAL DPro: Monitoring, Evaluation, Accountability and Learning for Development Professionals.* Catholic Relief Services, Humentum, and the Humanitarian Leadership Academy (Downloaded from http://mealdpro.org/ on July 25, 2019).

DataCamp. 2019. "Inferential Statistics Course." Accessed July 9, 2019. https://www.datacamp.com/community/open-courses/inferential-statistics.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. *Using Randomization in Development Economics Research: A Toolkit*. Vol. 6059.*Discussion Paper Series*. London: Centre for Economic Policy Research.

Frost, Jim. 2020. "Comparing Hypothesis Tests for Continuous, Binary, and Count Data." Statistics By Jim. Accessed 28 October 2020. https://statisticsbyjim.com/hypothesis-testing/comparing-hypothesis-tests-data-types/#:~:text=Additionally%2C%20the%20samples%20sizes%20are,sizes%20can%20become%20quite%20large.

Goldstein, Harvey, William Browne, and Jon Rasbash. 2002. "Partitioning Variation in Multilevel Models." *Understanding Statistics* 1 (4): 223-231. https://doi.org/10.1207/S15328031US0104_02.

Handa, Sudhanshu, Thomas de Hoop, Mitchell Morey, and David Seidenfeld. 2018. "ICC Values in International Development: Evidence across Many Domains in sub-Saharan Africa." Centre for the Study of African Economics conference, United Kingdom.

Hansen, Morris H., and Willimam N. Hurwitz. 1942. "On the Theory of Sampling from Finite Populations." *The Annals of Mathematical Statistics* 14 (4): 333-362. https://doi.org/https://www.jstor.org/stable/2235923.

Higgins, JPT, T Li, and JJ Deeks, eds. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*. Edited by JPT Higgins, J Thomas, J Chandler, M Cumpston, T Li, MJ Page and VA Welch. Version 6.0 ed, *Cochrane Handbook for Systematic Reviews of Interventions*: Available from www.handbook.cochrane.org.

Horvitz, Daniel G., and Donovan J. Thompson. 1952. "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association* 47 (260): 663-685. https://doi.org/10.2307/2280784.

Killip, Shersten, Ziyad Mahfoud, and Kevin Pearce. 2004. "What is an intracluster correlation coefficient? Crucial concepts for primary care researchers." *Annals of Family Medicine* 2 (3): 204-208.

Lana, Milza M. 2012. "The effects of line spacing and harvest time on processing yield and root size of carrot for Cenourete® production." *Horticultura Brasileira* 30 (2): 7.

Lee, Jin Young, and Gary Solon. 2011. "The fragility of estimated effects of unilateral divorce laws on divorce rates." *The BE Journal of Economic Analysis & Policy* 11 (1).

McConnell, Brendon, and Marcos Vera-Hernandez. 2015 2015. *Going beyond simple sample size calculations: a practitioner's guide.* Institute for Fiscal Studies.

Moss, Cami, Tesfaye Hailu Bekele, Mihretab Melesse Salasibew, Joanna Sturgess, Girmay Ayana, Desalegn Kuche, Solomon Eshetu, Andinet Abera, Elizabeth Allen, and Alan D Dangour. 2018. "Sustainable Undernutrition

Reduction in Ethiopia (SURE) evaluation study: a protocol to evaluate impact, process and context of a large-scale integrated health and agriculture programme to improve complementary feeding in Ethiopia." *BMJ Open* 8 (7). https://doi.org/10.1136/bmjopen-2018-022028.

National Research Council. 2002. *Studies of Welfare Populations: Data Collection and Research Issues*. Edited by Michele Ver Ploeg, Robert A. Moffitt and Constance F. Citro. Washington, DC: The National Academies Press.

Noggle, Eric. 2017. *The SILC Financial Diaries* Catholic Relief Services (Baltimore, MD).

Raab, Gillian, and Susan Purdon. 2009. "5.2.1 How the weights are calculated using population data." *Practical Exemplars on the Analysis of Surveys*. ReStore Project, National Centre for Research Methods (NCRM). Last Modified 5 July 2009. Accessed 22 April 2020. http://www.restore.ac.uk/PEAS/nonresponse.php.

Schochet, Peter Z. 2013. "Statistical Power for School-Based RCTs With Binary Outcomes." *Journal of Research on Educational Effectiveness* 6 (3): 263-294. https://doi.org/10.1080/19345747.2012.725803.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301-316. https://doi.org/10.3368/jhr.50.2.301.

Stukel, Diana Maria. 2018a. *Feed the Future Population-Based Survey Sampling Guide.* Food and Nutrition Technical Assistance Project, FHI 360 (Washington, DC. Downloaded from https://www.fantaproject.org/sites/default/files/resources/FTF-PBS-Sampling%20Guide-Apr2018.pdf on July 8, 2019: FHI 360).

---. 2018b. *Participant-Based Survey Sampling Guide for Feed the Future Annual Monitoring Indicators.* Food and Nutrition Technical Assistance Project, FHI 360 (Washington, DC. Downloaded from https://www.fantaproject.org/sites/default/files/resources/Sampling-Guide-Participant-Based-Surveys-Sep2018_0.pdf on July 8, 2019: FHI 360).

Sullivan, Lisa. 2019. "Power and Sample Size Determination." Accessed July 19, 2019. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Power/BS704_Power_print.html.

Technical Guidance for Monitoring, Evaluation, and Reporting for Emergency Activities. 2022. USAID Bureau for Humanitarian Assistance (Washington, DC. Downloaded from https://www.usaid.gov/sites/default/files/2022-05/BHA_Emergency_ME_Guidance_February_2022.pdf on December 18, 2023).

The Statistical Engineering Division. 1996. DATAPLOT Reference Manual. Downloaded from https://www.itl.nist.gov/div898/software/dataplot/refman2/ch2/weightsd.pdf on July 25, 2019: National Institutes of Standards and Technology, Information Technology Laboratory.

Thompson, Steven K. 2012. *Sampling*. Edited by Walter A. Shewhart and Samuel S. Wilks. Third ed.*Wiley Series in Probability and Statistics*. Hoboken, New Jersey: Wiley.

Thomsen, Ib, Dinke Tesfu, and David A. Binder. 1986. "Estimation of Design Effects and Intraclass Correlations When Using Outdated Measures of Size." *International Statistical Review / Revue Internationale De Statistique* 54 (3): 343-349. https://doi.org/10.2307/1403063