FANTA III

FOOD AND NUTRITION TECHNICAL ASSISTANCE

USAID FROM THE AMERICAN PEOPLE

Ensuring Attainment of Required Survey Sample Size of Children under 5 Years of Age through the Projection of the Appropriate Number of Households to Randomly Sample

Marcello Pagano, Harvard University
Diana Maria Stukel, FANTA, FHI 360

June 2018

FANTA
FHI 360
1825 Connecticut Ave., NW
Washington, DC 20009-5721
Tel: 202-884-8000 Fax: 202-884-8432
fantamail@fhi360.org www.fantaproject.org

fhi 360 THE SCIENCE OF IMPROVING LIVES

This report is made possible by the generous support of the American people through the support of the Office of Maternal and Child Health and Nutrition, Bureau for Global Health, and the Office of Food for Peace, U.S. Agency for International Development (USAID) and USAID Office of Food for Peace, under terms of Cooperative Agreement No. AID-OAA-A-12-00005, through the Food and Nutrition Technical Assistance III Project (FANTA), managed by FHI 360. The contents are the responsibility of FHI 360 and do not necessarily reflect the views of USAID or the United States Government.

June 2018

Recommended Citation

Pagano, Marcello and Stukel, Diana Maria. 2018. *Ensuring Attainment of Required Survey Sample Size of Children under 5 Years of Age through the Projection of the Appropriate Number of Households to Randomly Sample.* Washington, DC: FHI 360/FANTA.

Contact Information

Food and Nutrition Technical Assistance III Project (FANTA)
FHI 360
1825 Connecticut Avenue, NW
Washington, DC 20009-5721
T 202-884-8000
F 202-884-8432
fantamail@fhi360.org
www.fantaproject.org

Contents

Abbreviations and Acronyms

DFAP	development food assistance program
FANTA	Food and Nutrition Technical Assistance III Project
FFP	USAID Office of Food for Peace
USAID	U.S. Agency for International Development
WHO	World Health Organization

Abstract

Many multipurpose sample surveys seek to obtain estimates of indicators at the individual level (e.g., “Prevalence of Stunted Children under Five Years of Age”), and calculate required sample sizes at that level. Unfortunately, the lowest level of sampling may occur at the household level (assuming information on stunting is gathered on all eligible children under the age of 5 within sampled households.), complicating the correspondence between the number of households sampled and the number of children on which information is collected. The overall sample size that must be achieved for such surveys is therefore related to key indicators on which the surveys seek measurement (such as that related to stunting), and the overall sample size is pegged at the individual level. The task is to determine how many randomly chosen households to survey to generate a predetermined sample size of children under the age of 5 years who live in those households.

There are two problems that must be overcome under the above scenario. First, the number of children that reside in any particular household is unknown before the survey. Some households will have no eligible children, some will have one eligible child and some will have multiple such children. In relation to the first problem, it is unknown whether a given sampled household will yield an eligible child until the household is contacted. So, on one hand, the larger the number of households are sampled, the better the chance of finding a sufficient number of children. On the other hand, time, economics, and other considerations argue that the number of sampled households be made as small as possible. The second problem is that some chosen households will not respond, irrespective of whether there are eligible children in that household or not. The challenge is to manage the inherent uncertainty in these problems and, if possible, to improve on currently existing methods used to choose the appropriate number of households to sample.

This paper first describes the data that was used to guide a proposal for addressing this challenge, namely, a collection of 18 typical household surveys. The paper then introduces a novel method of approaching the problem by fitting a Poisson distribution. Subsequently, the data from the 18 surveys are also used to suggest a sampling distribution of the unknown parameters that are used to create a second statistical method to solve the problem at hand. This second proposed method uses confidence intervals, whose confidence levels are chosen by survey implementers to project appropriate household sample sizes; the greater the confidence, the larger the sample of households chosen. A comparison is made between the new proposed methods and the two existing methods. Finally, the paper provides a summary of the results in relation to all methods under consideration as well as recommendations for future use.

1. Introduction to the Problem

When designing a household survey where the key indicators that drive the overall sample size of the survey are at the individual child level (such as "Prevalence of Stunted Children under Five Years of Age"), the first step is to decide how big a sample of children under 5 years of age[1] to choose, assuming that the lowest level of random sampling occurs at the household and that all eligible children (rather than a subset of eligible children) within a sampled household have their anthropometric measurements taken. The next step is typically to determine how many households to randomly sample to ensure the required sample size of children. These two numbers—the required sample size of children (called n) and the number of households to randomly select (called N)—may not be the same for two reasons. The first reason relates to non-response, when the caregivers of some eligible children from whom data are to be collected within the sampled households cannot be reached or refuse to participate in the survey, and the second reason is that children under 5 may not be uniformly distributed throughout the sampled households. Indeed, it is unknown how many, or even if any, eligible children reside in a particular sampled household until an attempt is made to contact the household. This paper considers both of these problems together and proposes a limited solution to appropriately predict the number of households to sample in light of these constraints.

In an ideal scenario, to be able to predict the number of households to sample to ensure the required sample size of children under 5 years of age, it would be essential to have access to two important parameters at the survey planning stage that are unfortunately available only after survey completion. They are: the actual proportion of households that are sampled and participate in the survey (called γ_0) and the actual average number of children under 5 per household that participate in the survey (called λ_0).[2] If the values of these two parameters were known at the survey design stage, it would be straightforward to relate the number of children required with the number of houses to sample, that is, using $n = \lambda_0 \gamma_0 N$, and solving for N.

Unfortunately, these two quantities are unknown prior to survey implementation. This paper presents two approaches to address this challenge. The first approach is to proceed with *estimates* of both quantities: the estimated proportion of households that are sampled and participate in the survey (called γ) and the estimated average number of children under 5 per household that participate in the survey (called λ). In all likelihood, these estimates come from sources external to the survey, such as prior studies or censuses, and the number of randomly sampled households are based on these estimates. The better these estimates are, the closer the achieved yield of eligible children will be to the required yield. The second approach is to treat these estimates as realizations of a random mechanism, and to model the distribution producing these estimates.

[1] To simplify presentation, the example of children under 5 years of age is used throughout the paper, but the methods developed extend to other sub-populations as well.

[2] The latter quantity is treated as a single proportion, but it is actually calculated as the product of two numbers: the average number of eligible children per household and the proportion of those children that participate in the survey.

2. Existing Solutions and the Data on Which the Empirical Investigations Are Based

To make the discussion concrete, consider the results of the aforementioned 18 baseline multipurpose population-based surveys. The 18 surveys were funded by the U.S. Agency for International Development (USAID) Office of Food for Peace (FFP), in support of USAID/FFP Development Food Assistance Programs (DFAPs) undertaken by various nongovernmental organizations in a variety of developing countries. All surveys were implemented by ICF International, under contract to USAID/FFP. Among other data points, ICF International collected height and age information to support the production of stunting rates on children under the age of 5 years. Table 1 lists the 18 surveys, together with their identifying numeric labels that are used to reference them throughout this document.

Table 1. USAID/FFP baseline surveys, countries of origin, and DFAP implementing organizations

Country	Survey Number	DFAP Implementing Organization
Guatemala	1	Catholic Relief Services
	2	Save the Children
Uganda	3	Mercy Corps
	4	ACDI/VOCA
Niger	5	Save the Children
	6	Catholic Relief Services
	7	Mercy Corps
Zimbabwe	8	Cultivating New Frontiers in Agriculture
	9	World Vision
Haiti	10	CARE
Madagascar	11	Adventist Development and Relief Agency
	12	Catholic Relief Services
Burundi	13	Catholic Relief Services
Nepal	14	Save the Children
	15	Mercy Corps
Malawi	16	Project Concern International
	17	Catholic Relief Services
Mali	18	CARE

Table 2 contains information associated with these surveys. (Note that some information was available at the survey design stage and some became available only after the survey was completed.) The second column of the table provides the required sample size of children under the age of 5 years to be achieved at each of two time points (e.g., for a baseline survey and for an end-line survey) based on a statistical test of differences on indicators of proportions.[3] The next two columns (third and fourth) display estimates of the average number of eligible children in a household (λ) and the average household response rate (γ), respectively. Both parameters are estimated at the survey design stage from external information. These parameter estimates are used in computing the Stukel-Deitchler inflator to calculate the number of households to randomly sample, using the required sample size of children under 5 from the second column as input.[4] The number of households to randomly sample using the Stukel-Deitchler inflator is shown in column 5.[5] Columns 6, 7, and 8 show quantities available only after survey completion. Columns 6 and 7 reveal the actual average number of responding children in the households visited (λ_0) and the actual household response rates achieved (γ_0) in each of these surveys, respectively, while column 8 shows the actual number of children realized by the surveys. The minimum, mean, and maximum of each column are listed at the bottom of the table, to provide summary statistics.

The second column of Table 2, the required sample size of children under 5 years, are the required targets, which are sometimes missed with a sample that is too small or too large. Presumably, the risks are not symmetric around the targeted amount, and are typically survey dependent. This paper notes only the size of the under- or overestimation of the sample size targets and does not report on the relative merits of either error.

Comparing the last column to the second column shows that in all 18 of the surveys there were more children in the sample than required—on average, a surplus of 927 (2,405 − 1,478) children. One reason why the Stukel-Deitchler method of deciding how many households to sample (the method of inflation used for all 18 surveys) overshot the requirement is that the estimated lambda (λ) and gamma (γ) are not the same as the actual lambda (λ_0) and gamma (γ_0), respectively. This is an issue regardless of which inflation method is used (which will be illustrated later in the paper), although it affects different methods differently. It is the crux of the estimation problem that this paper addresses.

Note from Table 2 that the parameters λ_0 and γ_0 are calculated from the survey data and so are survey specific, yet their estimates (λ and λ) by country are obtained from external sources. Furthermore, the surveys across different DFAPs within a particular country all share the same λ but have different λ_0; similarly, they all share the same γ but have different γ_0. This reflects the unavailability of better parameter estimates at the granularity of a particular DFAP within a given country.

[3] The details of how these child-level sample sizes were derived is beyond the scope of this paper. For more information, see Stukel, Diana Maria. 2018. *Feed the Future Population-Based Survey Sampling Guide*, available at https://agrilinks.org/post/feed-future-zoi-survey-methods, hyperlinked under “2.1 Sampling manual.”

[4] A full description and derivation of the Stukel-Deitchler inflator can be found at: Stukel, Diana Maria. 2018. *Feed the Future Population-Based Survey Sampling Guide, Annex A*, available at https://agrilinks.org/post/feed-future-zoi-survey-methods, hyperlinked under “2.1 Sampling manual.”

[5] Note that each of these household level sample sizes in column 5 is a multiple of 30. That is because the survey used cluster sampling with clusters of size 30, and therefore the computed household level sample sizes were rounded up to the nearest multiple of 30.

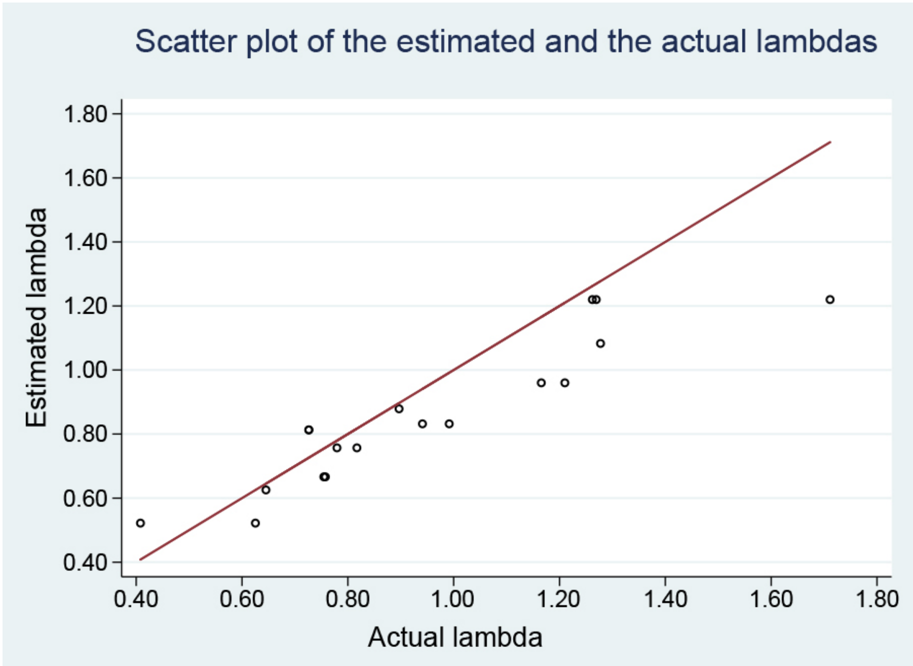
Table 2. USAID/FFP baseline surveys and their properties

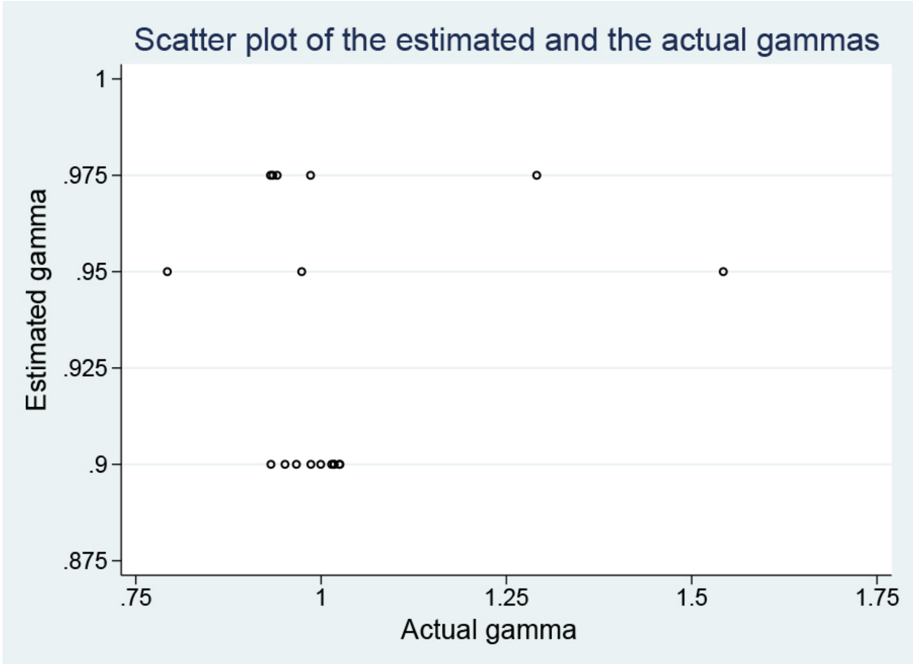
Survey Number	Pre-Survey Calculations				Post-Survey Calculations		
	Required Sample Size of Children under 5 years (n)	Estimated Lambda (λ)	Estimated Gamma (γ)	Households to Randomly Sample Based on Stukel-Deitchler Inflator	Actual Lambda (λ_0)	Actual Gamma (γ_0)	Actual Sample Size of Children under 5 years
1	1694	0.832	0.900	3000	0.991	1.025	3048
2	1694	0.832	0.900	3000	0.941	0.932	2632
3	1557	0.960	0.900	2400	1.210	1.000	2903
4	1557	0.960	0.900	2400	1.166	0.986	2759
5	1686	1.220	0.900	2400	1.711	1.018	4179
6	1686	1.220	0.900	2400	1.269	1.025	3124
7	1686	1.220	0.900	2400	1.262	1.014	3072
8	1227	0.667	0.900	2610	0.757	0.951	1881
9	1227	0.667	0.900	2610	0.754	0.967	1904
10	980	0.626	0.900	2220	0.651	1.007	1457
11	1454	0.757	0.975	2580	0.817	0.941	1983
12	1454	0.757	0.975	2580	0.779	0.935	1880
13	1434	0.879	0.975	2250	0.896	0.986	1989
14	1380	0.522	0.950	3420	0.408	0.954	1332
15	1380	0.522	0.950	3420	0.625	0.954	2041
16	1442	0.813	0.975	2400	0.902	0.995	2153
17	1442	0.813	0.975	2400	0.726	0.973	1695
18	1627	1.083	0.950	2220	1.278	0.974	2762
Minimum	**980**	**0.522**	**0.900**	**2220**	**0.408**	**0.932**	**1332**
Mean	**1478**	**0.85**	**0.929**	**2595**	**0.952**	**.980**	**2405**
Maximum	**1694**	**1.220**	**0.975**	**3420**	**1.711**	**1.025**	**4179**

Indeed, the discrepancies between column 3 and column 6 (the two lambdas) are large, as are the discrepancies between the two gammas (column 4 and column 7), as displayed in Figure 1. The horizontal variable for the graph of "actual gamma" may show a value greater than 1 (implying a household response rate of greater than 100%) because in some instances more households were visited than prescribed. The graphs are not meant to imply that it is possible to predict λ_0 and γ_0 perfectly. Indeed, the opposite is generally true. The challenge is to minimize the planned number of households to randomly sample and yet to achieve at least the required sample size of children while not overestimating to a great degree the number of households to sample.

Figure 1. Estimated and actual lambda values (left graph) and gamma values (right graph)

Note that the two graphs have different scales. The forty-five degree line is superimposed on the top graph.





3. Alternative Solution 1: The Poisson Method

Since lambda and gamma have multiplicative and sequential effects on the sample size, consider their product, kappa, $\kappa = \gamma * \lambda$. We model the number of households to sample, *N*, by assuming the average yield of children from each house is κ (the average number of eligible children per sampled and responding household), and by assuming data are collected on **all** eligible children within each sampled household (in contrast with selecting one eligible child at random).[6] If *X* denotes the number of eligible children in this sample of *N* households visited, then the mean of *X* is $N\kappa$ in this model. We can set the required sample size of children so that $n = N\kappa$, and solve for N. This method of setting the average yield to the required yield is called the Magnani solution[7] (even though the method did not originate with Magnani; it is ubiquitous in the literature[8]).

Alternatively, under the Poisson method, the probability structure of the Poisson Probability Distribution is used; the Poisson method uses a parameter, α, and under the Poisson distribution, the smallest *N* can be chosen so that:

$$\Pr(X \geq desired\ sample\ size\ of\ children) = 1 - \alpha$$

for a given confidence level $1 - \alpha$. This means that the survey will yield a sufficiently large sample with probability $(1 - \alpha)$.

Operationally, because of the large sample sizes needed, the "Normal" approximation to the Poisson distribution can be used in solving for *N*—and utilizing the fact that the mean and variance are the same for a Poisson variate (i.e., both equal to $N\kappa$). As before, it is assumed that the required sample size of children is n. Then, *N* can be solved for using the above equation. That is, the smallest N is chosen so that:

$$\Pr(X \geq n) = 1 - \alpha$$

$$Pr\left[\frac{X - N\kappa}{\sqrt{N\kappa}} \geq \frac{n - N\kappa}{\sqrt{N\kappa}}\right] = 1 - \alpha$$

$$Pr\left[Z \geq \frac{n - N\kappa}{\sqrt{N\kappa}}\right] = 1 - \alpha$$

where *Z* is a standard Normal variate. So, for example, if $\propto = 0.05$, thus setting the chances of approaching a sufficient number of houses to 0.95, then $Z = 1.645$. If $N\kappa > n$ is imposed according to the above equation, we have that:

$$\frac{N\kappa - n}{\sqrt{N\kappa}} \geq 1.645$$

[6] It is worth noting that the Poisson method hinges on the assumption of choosing **all** eligible children within a sampled household. In sample designs where **one** eligible child is randomly selected within a sampled household, it is no longer appropriate to use the Poisson method. In that case, a Binomial method, based on the Binomial distribution, is a more appropriate approach. The latter method is not presented here, but the authors can provide details on it upon request.

[7] See Magnani, Robert. 1999. *FANTA Sampling Guide*. Washington, DC: FHI 360.

[8] For example, this is the method promoted by DHS. See: ICF International. 2012. *Demographic and Health Survey Sampling and Household Listing Manual.* Page 11. MEASURE DHS, Calverton, Maryland: ICF International, Page 11. The method is also promoted by the World Health Organization (WHO) in: WHO Library Cataloguing in Publication Data. 2007. *Assessing tuberculosis prevalence through population-based surveys* (ISBN 978 92 9061 314 5).

Replacing the 1.645 by 2.326 yields the solution for the case when $\propto = 0.01$, in which case the chances of approaching a sufficient number of households increase to 0.99. Assuming $\propto = 0.05$, this equation can be solved by squaring both sides and solving the resultant quadratic equation (using the larger of the two quadratic solutions, since it satisfies the requirement that $N\kappa > \text{n}$):

$$N \geq \frac{(b + \sqrt{b^2 - 4n^2})}{2\kappa} \qquad (1)$$

In Equation 1, $b = 2n + \ 1.645^2$ for the 95% solution. Replace 1.645 by 2.326 for the 90% solution and by 1.282 for the 80% solution. For the 50% solution, replace 1.645 with 0, which results in $N = n / \kappa$; that is, the formula reduces to the Magnani solution. Finally, when the sampling design is a clustered sample with the clusters of size 30, *N* should be rounded up to a multiple of 30.

Table 3 provides the results of the Poisson method applied to the 18 surveys shown in Table 2, with the number of households to sample rounded up to the closest multiple of 30. For comparison purposes, the last column includes the number of households to sample based on the Stukel-Deitchler method. In this table, it is assumed that there is no error in the estimation of kappa, and therefore the actual value of kappa from Table 2 is used in the computation of all methods, in an attempt to investigate how the various methods stack up against each other based on theoretical considerations only.

Table 3. The number of households (*N*) to sample for three methods (Magnani, Poisson, Stukel-Deitchler) using the actual kappa (κ_0) as an input parameter

Survey Number	Actual Kappa	Poisson Method				Stukel-Deitchler
		50% (Magnani)	80%	90%	95%	
1	1.016	1680	1740	1740	1770	2310
2	0.8773	1950	2010	2010	2040	1710
3	1.2096	1290	1350	1350	1380	1770
4	1.1496	1380	1410	1440	1440	1920
5	1.7413	990	1020	1020	1020	1350
6	1.3017	1320	1350	1350	1380	1590
7	1.28	1320	1380	1380	1410	1650
8	0.7207	1710	1770	1800	1830	2220
9	0.7295	1710	1770	1770	1800	2190
10	0.6563	1500	1560	1590	1590	1890
11	0.7686	1920	1980	1980	2010	2430
12	0.7287	2010	2070	2100	2130	2580
13	0.884	1650	1680	1710	1710	2160
14	0.3895	3570	3690	3720	3750	4140
15	0.5968	2340	2400	2430	2460	2850
16	0.8971	1620	1680	1680	1710	2190
17	0.7063	2070	2130	2160	2160	2610
18	1.2441	1320	1350	1380	1380	1890
Mean		**1742**	**1797**	**1812**	**1832**	**2192**

In Table 3, the reason why the projected *N* is sometimes the same for different confidence levels—for example, in Survey 1, 1740 households yield confidences of both 80% and 90%—is that all these *N*s are rounded up to be multiples of 30 and some of them are close in value prior to rounding.

To emulate the yield of eligible children that would have resulted by sampling the number of households in Table 3, the values in Table 3 are multiplied by the actual kappa, κ_0. The results are shown in Table 4.

Table 4. The emulated yield of children in the sample calculated by assuming the number of households in Table 3 under the three methods (Magnani, Poisson, Stukel-Deitchler) and using the actual kappa (κ_0) as an input parameter

Survey Number	Required Sample of Children (*n*)	Actual Kappa	Poisson Method				Stukel-Deitchler
			50% (Magnani)	80%	90%	95%	
1	1694	1.016	1706.9	1767.8	1767.8	1798.3	2347.0
2	1694	0.8773	1710.8	1763.4	1763.4	1789.8	1500.2
3	1557	1.2096	1560.4	1632.9	1632.9	1669.2	2141.0
4	1557	1.1496	1586.4	1620.9	1655.4	1655.4	2207.2
5	1686	1.7413	1723.8	1776.1	1776.1	1776.1	2350.7
6	1686	1.3017	1718.2	1757.2	1757.2	1796.3	2069.6
7	1686	1.2800	1689.6	1766.4	1766.4	1804.8	2112.0
8	1227	0.7207	1232.4	1275.6	1297.2	1318.9	1599.9
9	1227	0.7295	1247.4	1291.2	1291.2	1313.1	1597.6
10	980	0.6563	984.5	1023.8	1043.5	1043.5	1240.4
11	1454	0.7686	1475.7	1521.8	1521.8	1544.9	1867.7
12	1454	0.7287	1464.7	1508.4	1530.2	1552.1	1880.0
13	1434	0.884	1458.6	1485.1	1511.6	1511.6	1909.4
14	1380	0.3895	1390.4	1437.2	1448.8	1460.5	1612.4
15	1380	0.5968	1396.5	1432.3	1450.2	1468.1	1700.8
16	1442	0.8971	1453.3	1507.1	1507.1	1534.0	1964.6
17	1442	0.7063	1461.9	1504.3	1525.5	1525.5	1843.3
18	1627	1.2441	1642.3	1679.6	1716.9	1716.9	2351.4
Mean	**1478**		**1494.7**	**1541.7**	**1553.5**	**1571.1**	**1905.3**

None of the projected sample sizes fall below the required sample size, *n*. Furthermore, the sample size projected by the Magnani method tracks the required sample size fairly closely, which bodes well for this method from a theoretical standpoint. As is to be expected, the Poisson method gives sample size values that increase from the 50% confidence (smallest) to the 95% confidence (largest). In Survey 1, to increase from 50% confidence to 95% confidence requires, on average, only an extra 90 households (1770 − 1680 = 90), as can be seen from Table 2. In the last column of Table 3, the Stukel-Deitchler method provides sample sizes that are greater than even the 95% confidence column; this is true of every survey except the second.

Tables 3 and 4 investigate how many households to sample and what is the yield of eligible children if the actual kappa (κ_0) were known at the design stage, when the sample sizes are computed. Of course, in practice, the actual kappa is not known at the design stage and the estimated kappa (κ) must be used instead. The paper now turns to studying the impact on sample size projections using the estimated kappa.

The success of these household-level sample size projections is very much dependent on how close the estimated κ is to the actual κ_0. To quantify the impact of estimating kappa on these household projections, it is possible to emulate what would have happened had the only available information prior to survey work (i.e., the estimated kappa, κ) been used instead. To do so, the number of households to sample in Table 3 is recomputed using the estimated kappa (κ). These calculations are shown in Table 5.

Table 5. The number of households (*N*) to sample for three methods (Magnani, Poisson, Stukel-Deitchler) using the estimated kappa (κ) as an input parameter

Survey Number	Estimated Kappa	Poisson Method				Stukel-Deitchler
		50%	80%	90%	95%	
1	0.75	2280	2340	2370	2400	3000
2	0.75	2280	2340	2370	2400	3000
3	0.86	1830	1890	1890	1920	2400
4	0.86	1830	1890	1890	1920	2400
5	1.10	1560	1590	1590	1620	2400
6	1.10	1560	1590	1590	1620	2400
7	1.10	1560	1590	1590	1620	2400
8	0.60	2070	2130	2160	2190	2610
9	0.60	2070	2130	2160	2190	2610
10	0.56	1770	1830	1860	1860	2220
11	0.74	1980	2040	2070	2100	2580
12	0.74	1980	2040	2070	2100	2580
13	0.86	1680	1740	1770	1770	2236
14	0.50	2790	2910	2910	2940	3420
15	0.50	2790	2910	2910	2940	3420
16	0.79	1830	1890	1920	1920	2400
17	0.79	1830	1890	1920	1920	2400
18	1.03	1590	1620	1650	1650	2220
Mean	**0.7905**	**1960**	**2020**	**2038**	**2060**	**2594**

In a manner similar to how Table 4 was generated, to emulate the yield of eligible children that would have resulted by sampling the number of households in Table 5, the values in Table 5 are multiplied by the estimated kappa, κ. The results are shown in Table 6.

For comparison, the last column shows the actual yields from the 18 surveys (as in Table 2). This column is labeled "Stukel-Deitchler" because the computation of the number of households to sample for the 18 surveys was based on the Stukel-Deitchler inflator prior to fieldwork, and the yield of children under 5 years of age realized from the fieldwork of these 18 surveys is equivalent to multiplying the last column of Table 5 by the estimated kappa, κ, making the comparison between the Poisson and Stukel-Deitchler methods equitable in this table.

Table 6. The emulated yield of children in the sample calculated by assuming the number of households in Table 5 under the three methods (Magnani, Poisson, Stukel-Deitchler), using the estimated kappa (κ) as an input parameter, and by multiplying the results in Table 5 by the actual kappa (κ_0)

Survey Number	Required Sample of Children (n)	Kappa Actual	Poisson Method				Stukel-Deitchler
			50%	80%	90%	95%	
1	1694	1.0160	2316.5	2377.4	2407.9	2438.4	3048
2	1694	0.8773	2000.3	2053.0	2079.3	2105.6	2632
3	1557	1.2096	2213.5	2286.1	2286.1	2322.4	2903
4	1557	1.1496	2103.7	2172.7	2172.7	2207.2	2759
5	1686	1.7413	2717.2	2769.5	2769.5	2821.7	4179
6	1686	1.3017	2029.1	2068.2	2068.2	2107.2	3124
7	1686	1.2800	1996.3	2034.7	2034.7	2073.1	3072
8	1227	0.7207	1491.8	1535.1	1556.7	1578.3	1881
9	1227	0.7295	1510.1	1553.8	1575.7	1597.6	1904
10	980	0.6563	1161.7	1201.0	1220.7	1220.7	1457
11	1454	0.7686	1521.8	1568.0	1591.0	1614.1	1983
12	1454	0.7287	1442.8	1486.5	1508.4	1530.2	1880
13	1434	0.884	1485.1	1538.2	1564.7	1564.7	1977
14	1380	0.3895	1086.6	1133.4	1133.4	1145.1	1332
15	1380	0.5968	1665.0	1736.6	1736.6	1754.5	2041
16	1442	0.8971	1641.7	1695.5	1722.4	1722.4	2153
17	1442	0.7063	1292.4	1334.8	1356.0	1356.0	1695
18	1627	1.2441	1979.2	2016.5	2053.9	2053.9	2762
Mean	**1478**	**0.9380**	**1759**	**1809**	**1824**	**1845**	**2377**
Average Excess Sample Size over Required			**281**	**331**	**346**	**367**	**899**

As expected, the values produced by the Poisson method are ordered from the smallest (50% confidence) to the largest (95% confidence). Over the 18 surveys, the sample sizes average 1759, 1809, 1824, and 1845, at 50% confidence, 80% confidence, 90% confidence, and 95% confidence, respectively, all smaller than the number of children actually surveyed under the Stukel-Deitchler method (the last column), which averaged 2377. The average excess sample size over that required (given in column 2) ranges between 281 children (Magnani method) and 899 children (Stukel-Deitchler method).

Table 7 shows the estimated kappa in column 3, which can be contrasted with the actual kappa in column 4. The relative error between them is computed in column 5. The remaining columns express the sample size produced by each of the methods under consideration as a percentage of excess/shortfall relative to the required sample size of children (n) in column 2.

Table 7. The emulated yield of children under 5 from Table 6 associated with the number of households in Table 5 expressed as relative percentage excess/shortfall of required sample size of children

Survey Number	Required Sample Size of Children (n)	Estimated Kappa (κ)	Actual Kappa (κ_0)	% Relative Error $(\kappa - \kappa_0) / \kappa_0$	Poisson Method				Stukel-Deitchler
					50% (Magnani)	80%	90%	95%	
1	1694	0.7488	1.0160	−26.3	36.7	40.3	42.1	43.9	79.9
2	1694	0.7488	0.8773	−14.7	18.1	21.2	22.7	24.3	55.4
3	1557	0.8640	1.2096	−28.6	42.2	46.8	46.8	49.2	86.4
4	1557	0.8640	1.1496	−24.8	35.1	39.5	39.5	41.8	77.2
5	1686	1.1000	1.7413	−36.8	61.2	64.3	64.3	67.4	147.9
6	1686	1.1000	1.3017	−16.7	20.3	22.7	22.7	25.0	85.3
7	1686	1.1000	1.2800	−14.1	18.4	20.7	20.7	23.0	82.2
8	1227	0.5998	0.7207	−16.8	21.6	25.1	26.9	28.6	53.3
9	1227	0.5998	0.7295	−17.8	23.1	26.6	28.4	30.2	55.2
10	980	0.5630	0.6563	−14.2	18.5	22.6	24.6	24.6	48.7
11	1454	0.7380	0.7686	−4.0	4.7	7.8	9.4	11.0	36.4
12	1454	0.7380	0.7287	1.3	−0.8	2.2	3.7	5.2	29.3
13	1434	0.8569	0.8840	−3.1	3.6	7.3	9.1	9.1	37.8
14	1380	0.4959	0.3895	27.3	−21.3	−17.9	−17.9	−17.0	−3.5
15	1380	0.4959	0.5968	−16.9	20.7	25.8	25.8	27.1	47.9
16	1442	0.7928	0.8971	−11.6	13.8	17.6	19.4	19.4	49.3
17	1442	0.7928	0.7063	12.3	−10.4	−7.4	−6.0	−6.0	17.5
18	1627	1.0300	1.2441	−17.2	21.6	23.9	26.2	26.2	69.8
Mean	**1478.2**	**0.7905**	**0.9387**	**−12.4**	**18.2**	**21.6**	**22.7**	**24.1**	**58.7**

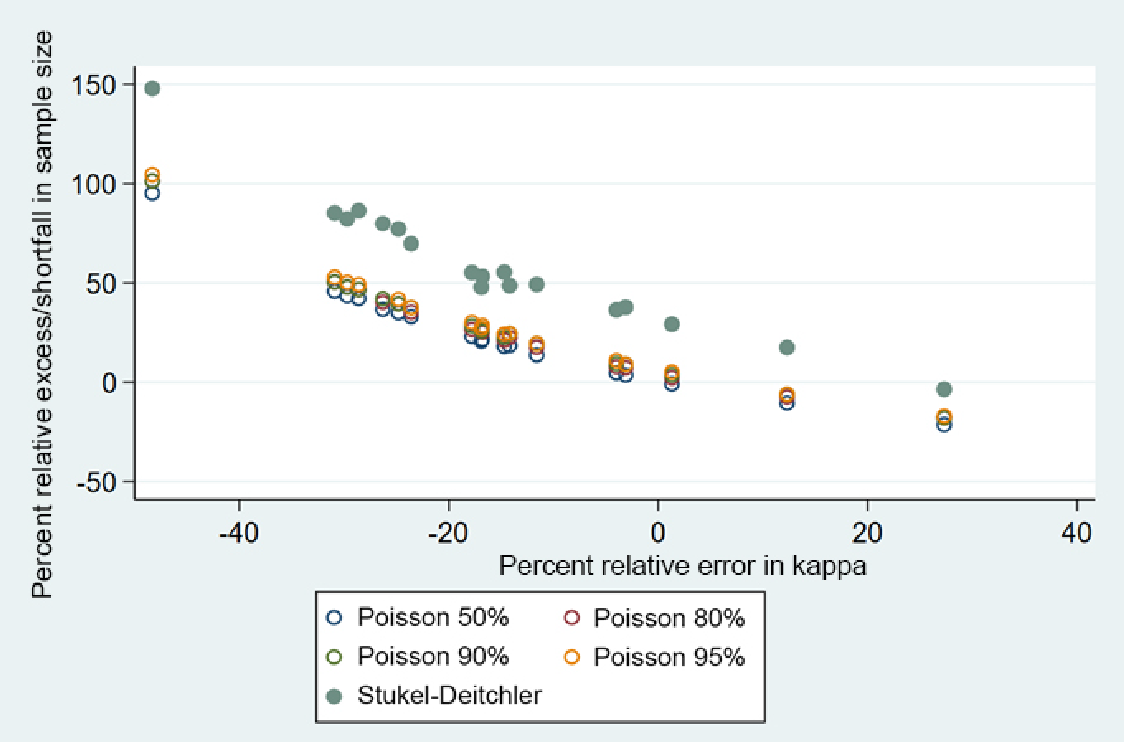
Note: Numbers in red indicate shortfalls.

The sample size based on the Magnani method fell short in Surveys 12, 14, and 17, and based on the Poisson method fell short in Surveys 14 and 17 for all confidence levels. The sample size based on the Stukel-Deitchler method fell short in Survey 14 only.

It is interesting to note that the sample sizes produced by the Magnani method did **not** fall short of the required sample size in roughly 50% of the surveys, as expected assuming the model on which its defining formula is based is correct. In fact, the Magnani method did not fall short for 15 out of 18 surveys. In these 15 surveys, the actual kappa is larger than the estimated one, so, in these cases, kappa is underestimated. Since for the Poisson method (of which Magnani is a special case) the number of households to sample (N) is inversely related to kappa (see Equation (1)), an underestimation of kappa tends to make the samples larger than necessary (for all confidence levels). For the Magnani method, although in principle one would expect 50% of the surveys to fall short of the required sample size and 50% of the surveys to exceed the required sample size, in fact this happens only because it is expected that kappa is overestimated for 50% of the surveys and underestimated for 50% of the surveys – and this did not happen for the 18 surveys considered. That only 3 of the 18 surveys (instead of 9 of the 18 surveys) have kappa values where the estimated kappa is larger than the actual kappa (i.e., kappa is overestimated), is worrisome. The comparison presented above is possibly confounding how sensitive each method is to the choice of estimated kappa. [9]

Figure 2 depicts Table 7 in a graphical format, with column 5 (% relative error in kappa) plotted on the horizontal axis and columns 6, 7, 8, 9, and 10 (percent relative excess/shortfall in sample size using Poisson and Stukel-Deitchler methods) plotted on the vertical axis.

Figure 2. Percent relative excess/shortfall in sample size in various methods vs. percent relative error in kappa



It can be seen from this graph that the ideal scenario (i.e., 0% overshoot/undershoot of sample size) is one where κ is close to κ_0 or where $(\kappa - \kappa_0)/\,\kappa_0$ is close to zero. This, of course, supports Table 3, which sets $\kappa = \kappa_0$. If only those surveys where, for example, $(\kappa - \kappa_0)/\,\kappa_0 < 20\%$ are considered, then the focus is somewhat shifted towards only those cases where kappa is relatively well estimated. Table 8 lists the 10 surveys for which this inequality is satisfied. That is, Table 8 is a subset of Table 7. Note that now all the kappa values are less than 1.

[9] The fact that the Magnani method overestimates the sample yield of children more often than it underestimates the sample yield of children may be exacerbated by the idiosyncratic rounding up to the nearest multiple of 30 of the number of households to sample – in order to accommodate cluster samples sizes of 30. This upward rounding may be according the Magnani method an additional cushion of sample size that would otherwise not be there. Removing the rounding may in fact bring us somewhat closer to a 50-50 split with regards to overestimation and underestimation.

Table 8. Emulated yield of children under 5 associated with the number of households in Table 5, restricted to relative error in kappa less than 20% (Subset of Table 7)

Survey Number	Required Sample Size of Children (n)	Estimated Kappa (κ)	Actual Kappa (κ_0)	% Relative Error $(\kappa - \kappa_0) \div \kappa_0$	Poisson Method				Stukel-Deitchler
					50% (Magnani)	80%	90%	95%	
2	1694	0.7488	0.8773	−14.7	18.1	21.2	22.7	24.3	55.4
8	1227	0.5998	0.7207	−16.8	21.6	25.1	26.9	28.6	53.3
9	1227	0.5998	0.7295	−17.8	23.1	26.6	28.4	30.2	55.2
10	980	0.5630	0.6563	−14.2	18.5	22.6	24.6	24.6	48.7
11	1454	0.7380	0.7686	−4.0	4.7	7.8	9.4	11.0	36.4
12	1454	0.7380	0.7287	1.3	−0.8	2.2	3.7	5.2	29.3
13	1434	0.8569	0.8840	−3.1	3.6	7.3	9.1	9.1	37.8
15	1380	0.4959	0.5968	−16.9	20.7	25.8	25.8	27.1	47.9
16	1442	0.7928	0.8971	−11.6	13.8	17.6	19.4	19.4	49.3
17	1442	0.7928	0.7063	12.3	−10.4	−7.4	−6.0	−6.0	17.5

Note: Numbers in red indicate shortfalls.

Note, again, that the Magnani method yields sample sizes that are smaller than required (Surveys 12 and 17) when the estimated kappa values (κ) are greater than the actual kappa values (κ_0). The Poisson method with confidence levels greater than 50% provide a small cushion over the Magnani solution. This cushion is sufficient enough to overcome the shortfall in Survey 12, but not in Survey 17. The reason for the insufficiency of the cushion is that in Survey 17 the relative discrepancy between the two kappa values is 12.3%; in other words, the estimated kappa is evidently much too different from the actual kappa in this case.

In summary, the Magnani method is inversely proportional to the estimated kappa and the performance of the method relates to how closely it tracks the actual kappa: If the estimated kappa is larger than the actual kappa, the resultant sample size is smaller than the required sample size of children; if the estimated kappa is smaller than the actual kappa, the resultant sample size is larger than the required sample size of children. The Poisson method at confidence levels greater than 50% yields slightly larger samples than the Magnani solution by design, and thus they build in a slight cushion against underestimating the sample size. The Stukel-Deitchler method yields sample sizes that are larger than the Poisson methods for all confidence levels and, in all but one survey, yield much larger sample sizes than are necessary.

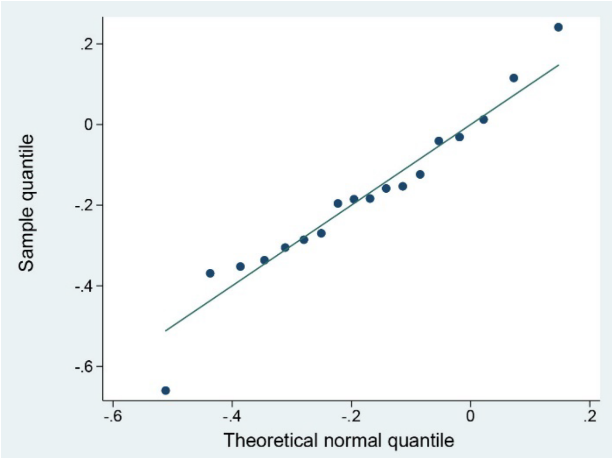
4. Alternative Solution 2: The Kappa Prediction Method

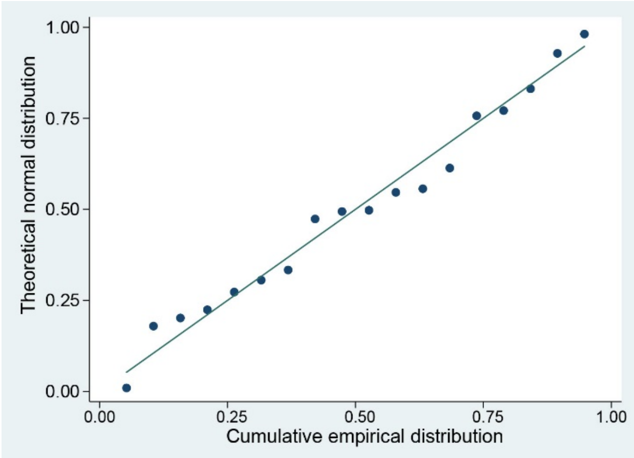
In the three methods contrasted to this point—Magnani, Poisson, and Stukel-Deitchler—the value of kappa is assumed to be a constant, both at the design of the survey (where the estimated kappa is used) and after the survey is completed (where the actual kappa is realized). None of the methods prescribes a principled approach for building in some insurance to cover prediction inaccuracies in the value of kappa. That kappa is often inaccurately estimated is to be expected, and the estimate available, κ (column 3 in Table 7), versus the observed, κ_0 (column 4 in Table 7), shows this clearly.

Another approach is to view the issue as a prediction problem, and to treat κ as a random variable with a distribution that can be used to predict its behavior. To devise a methodology for the determination of the number of households to sample, the distribution of κ around κ_0 needs to be studied. The data from the 18 USAID/FFP baseline surveys provides an empirical basis to help initiate an investigation into the behavior of the distribution of κ.

Inasmuch as this sample of 18 USAID/FFP baseline surveys is representative of such surveys of its ilk, consider the distribution of the logarithm of the ratio κ / κ_0 over the 18 surveys. Ideally, this ratio is 1 and would have no variance. In reality, however, if Q–Q and P–P (Normal) plots of the logarithm of this ratio are graphed, the variability of the ratio is evident (Figure 3).

Figure 3. Q–Q plot for log(κ/κ_0) on the left and P–P plot for log(κ/κ_0) on the right





These plots help determine whether or not log(κ/κ_0) follows a Normal distribution. For the Q–Q plot, the non-cumulative distribution of log(κ/κ_0) is plotted against the theoretical non-cumulative Normal distribution. For the P–P plot, the cumulative distribution of log(κ/κ_0) is plotted against the theoretical cumulative Normal distribution. Overall, if the log(κ/κ_0) follows a Normal distribution, then the pattern of data points from these surveys would be expected to follow a straight line in both plots. In principle, the Q–Q plot magnifies deviations from normality in the tails, whereas the P–P plot magnifies deviations from normality in the center of the distribution. Judging from the agreement with a straight line in these two plots, they support treating log(κ/κ_0) as following a Normal distribution. (The outlier in the Q–Q plot is Survey 5, which had the highest negative percent relative error in Table 7.) The empirical mean (of the log transform) in these 18 surveys is −0.18 with a standard deviation of 0.15. This is approximated as a mean zero distribution because −0.18 is so close to zero. This agrees with the notion that estimates of κ_0 tend to be "unbiased"—that is to say, typical estimates are just as likely to overestimate as to underestimate the size of κ_0.

Utilizing a distribution for this ratio, a confidence interval around κ_0 can be determined to guide the choice of the number of households to sample. Different distributions will, of course, provide different confidence intervals. This discussion uses the Lognormal distribution to illustrate this method.

All the algorithms use as input the required sample size of children under 5 (n) and the estimated kappa, and provide as output the number of households to sample, N. If κ_0 were known, the number of households to sample to obtain the required sample size of children would also be known ($N = n / \kappa_0$ under the Magnani method.) However, the actual value of kappa, κ_0, is not known and therefore the problem is framed as one of estimating the parameter κ_0 by predicting the value it will take using its underlying probability distribution.

Because the actual value of kappa (using results from the survey) is obtained through fieldwork, it is possible to judge how well the proposed method works in settings where the results of such surveys are available, in this case, from the 18 USAID/FFP baseline surveys referenced earlier. If kappa is overestimated, too few households are sampled and the sample size falls short of the required number of children. If kappa is underestimated, too many households are sampled and there is a surplus in relation to the required number of children. Of course, it is preferable to end up with more, rather than fewer, children than required. This argues against a point estimate, and for a one-sided confidence interval for κ_0.

For the Kappa Prediction method, the desired confidence level α is chosen. The interval for κ_0 is then determined so that, using the estimate κ and Z_α (the αth percentile of the standard Normal distribution[10]), we have that:

$$\Pr\left(\frac{\ln(\kappa/\kappa_0)}{0.15} \leq Z_\alpha\right) = 1 - \alpha \qquad (2)$$

The α is chosen to reflect the gravity of the consequence of underestimating versus overestimating the sample size. For example, if the preference is equal with regard to underestimation versus overestimation, $\alpha = 0.5$ (i.e., 50%) is chosen so that $Z_\alpha = 0$.

The confidence interval associated with Equation (2) provides a range of values of possible κ_0 and a sample size that is too small for any actual value of κ_0 within this interval can be guarded against by using the smallest κ_0 in the interval, called the "estimated actual kappa" and denoted $\widehat{\kappa_0}$. Thus:

$$\widehat{\kappa_0} = \kappa e^{-0.15 Z_\alpha} \qquad (3)$$

With the choice of $\widehat{\kappa_0}$ shown in Equation (3), the appropriate number of households to sample can be calculated (in multiples of 30) to achieve the required number of children as follows[11]:

$$N = \left\lceil \frac{n}{\kappa e^{-0.15 Z_\alpha}} \right\rceil_{30}^{+} \qquad (4)$$

The square bracket notation in Equation (4) refers to rounding the term inside the bracket to the next highest multiple of 30. Note that this method reduces to the Magnani method when $\alpha = 0.5$, which can be seen by setting $Z_\alpha = 0$, the 50th percentile of the standard Normal distribution, in Equation (4).[12]

[10] For example, if $\alpha = 0.05$, then $Z_\alpha = 1.645$.

[11] If cluster samples of size 30 are not being used, Equation (4) should be modified appropriately to reflect the appropriate cluster sample size.

[12] Note that Equation (4) uses the specific value of 0.15 for the standard deviation of $\log(\kappa/\kappa_0)$, which is derived using the data from the 18 USAID/FFP baseline surveys. For the Kappa Prediction method to be more broadly generalizable to surveys outside these 18 surveys, a value for the standard deviation of $\log(\kappa/\kappa_0)$ that is consistent with the survey(s) on which the method is being applied is required. Survey implementers should look to a spectrum of past surveys in similar countries using similar populations to obtain values for κ and κ_0 to estimate the standard deviation of $\log(\kappa/\kappa_0)$. The potential unavailability of such values for this standard deviation from prior surveys could pose limitations on the ability to implement this method in practice.

If this formula is applied to the 18 surveys with, in turn, $\alpha = 50\%, 20\%, 10\%, and\ 5\%$, (or $1 - \alpha = 50\%, 80\%, 90\%, and\ 95\%$), the results shown in Table 9 are obtained.

Table 9. Number of households to sample and expected yield of children under 5 for various confidence levels using the Kappa Prediction method

Survey Number	Required Sample Size of Children (n)	Estimated Kappa (κ)	Smallest Number of Households to Sample Based on Kappa Prediction				Actual Kappa (κ_0)	Expected Yield of Children under 5 Years of Age Based on Kappa Prediction			
			50%	80%	90%	95%		50%	80%	90%	95%
1	1694	0.75	2280	2580	2760	2910	1.02	2316.5	2621.3	2804.2	2956.6
2	1694	0.75	2280	2580	2760	2910	0.88	2000.3	2263.5	2421.4	2553.0
3	1557	0.86	1830	2070	2190	2310	1.21	2213.5	2503.8	2649.0	2794.1
4	1557	0.86	1830	2070	2190	2310	1.15	2103.7	2379.6	2517.6	2655.5
5	1686	1.1	1560	1770	1890	1980	1.74	2716.4	3082.0	3291.0	3447.7
6	1686	1.1	1560	1770	1890	1980	1.30	2030.6	2303.9	2460.1	2577.3
7	1686	1.1	1560	1770	1890	1980	1.28	1996.8	2265.6	2419.2	2534.4
8	1227	0.60	2070	2340	2490	2640	0.72	1491.8	1686.4	1794.5	1902.6
9	1227	0.60	2070	2340	2490	2640	0.73	1510.1	1707.0	1816.5	1925.9
10	980	0.56	1770	1980	2130	2250	0.66	1161.7	1299.5	1397.9	1476.7
11	1454	0.74	1980	2250	2400	2550	0.77	1521.8	1729.4	1844.7	1959.9
12	1454	0.74	1980	2250	2400	2550	0.73	1442.8	1639.5	1748.8	1858.1
13	1434	0.86	1680	1920	2040	2160	0.88	1485.1	1697.3	1803.4	1909.4
14	1380	0.50	2790	3180	3390	3570	0.39	1086.4	1238.3	1320.0	1390.1
15	1380	0.50	2790	3180	3390	3570	0.60	1664.2	1896.9	2022.1	2129.5
16	1442	0.79	1830	2070	2220	2340	0.90	1641.7	1857.1	1991.6	2099.3
17	1442	0.79	1830	2070	2220	2340	0.71	1292.0	1461.5	1567.4	1652.1
18	1627	1.03	1590	1800	1920	2040	1.24	1979.0	2240.4	2389.7	2539.1

The second column shows the required sample size of children under the age of 5 (n). The third column shows the estimated kappa (κ), which is used as an input parameter in Equation (3). Equation (3), in turn, is used as an input to Equation (4), which is used to compute the next four columns. Columns 4, 5, 6, and 7 are the number of households under the Kappa Prediction method, calculated using the confidence levels of 50%, 80%, 90%, and 95%, respectively. The next column (column 8) gives the actual kappa (κ_0) achieved for each survey using the original data (from Table 7). This is used to calculate the next four columns—one each for each of the confidence levels—to obtain the expected yield of children under 5 years of age resulting from sampling the number of households in columns 4, 5, 6, and 7. The numbers in red are those instances when either the expected sample sizes of children in columns 9, 10, 11, and 12 are smaller than the required number of children in column 2, or equivalently, the estimated kappa (κ) in column 3 is larger than the actual kappa (κ_0) in column 8.

Below are summarized the behaviors of the proposed Kappa Prediction method for these 18 surveys using the results from Table 9:

1. It can be seen from the 50% confidence level column in Table 9 that 15 surveys overestimated the sample sizes of children and that 3 surveys underestimated the sample sizes of children. That this is not the expected 50-50 split (that is, 9 surveys overestimating and 9 surveys underestimating) is due to the fact that 7 of the 18 actual kappa values (κ_0) are larger than 1.
2. The 80% and 90% confidence level columns in Table 9 show that only one survey (Survey 14) underestimated the sample sizes of children. Once again, this underestimation is probably related to the value of κ_0. However, given the modest cushion provided over the 50% confidence level solution at the 80% and 90% levels, these might be more attractive choices for the confidence level to use for these particular surveys.
3. The 95% confidence level column in Table 9 shows that there is no underestimation. This is not, however, to say that overestimation is without cost. This is explored further below.

An analysis of the values highlighted in red in Table 9 provides further insights. Consider Table 10, generated from the 18 USAID/FFP baseline surveys using values from Table 7.

Table 10. Ranking across surveys of ratio (κ / κ_0)

Survey	Estimated Kappa (κ)	Actual Kappa (κ_0)	Ratio (κ/κ_0)	Ranking of Ratio across Surveys
1	0.749	1.016	0.737	16
2	0.749	0.877	0.854	9
3	0.864	1.210	0.714	17
4	0.864	1.150	0.751	15
5	1.100	1.741	0.632	18
6	1.100	1.302	0.845	10
7	1.100	1.280	0.859	7
8	0.600	0.721	0.832	11
9	0.600	0.730	0.822	14
10	0.563	0.656	0.858	8
11	0.738	0.769	0.960	5
12	0.738	0.729	1.012	3
13	0.857	0.884	0.969	4
14	0.496	0.389	1.275	1
15	0.496	0.597	0.831	12
16	0.793	0.897	0.884	6
17	0.793	0.706	1.123	2
18	1.030	1.244	0.828	13

This table shows the relationship between the estimated kappa and the actual kappa for the 18 surveys. The fourth column shows the ratio of these two quantities, and the last column shows the rank of the size of the ratio across the 18 surveys, with 1 being the rank associated with the highest value of the ratio. The surveys that have values highlighted in red in the 50% column in Table 9 (Surveys 12, 14, and 17, where the expected number of children under 5 years is underestimated relative to the required number of children) are ranked 3, 1, and 2, respectively, in this table. These results are not surprising, because these are the only values of the ratio that exceed 1, but they also serve as motivation to make this ratio as close to 1 as possible, noting that those values of the ratio above 1 tend to underestimate the required sample size of children and those values of the ratio less than 1 tend to overestimate the required sample size of children.

5. Comparison of the Methods

Table 11 provides the number of households to sample according to four methods: the Magnani method, the Poisson method for three confidence levels (80%, 90%, and 95%), the Kappa Prediction method for three confidence levels (80%, 90%, and 95%), and the Stukel-Deitchler method. The Magnani column is common to both the Poisson method at the 50% confidence level and the Kappa Prediction method at the 50% confidence level.

Table 11. Number of households to sample using the four methods (Magnani, Poisson, Kappa Prediction, Stukel-Deitchler)

Survey Number	Magnani	Poisson			Kappa Prediction			Stukel-Deitchler
	50%	80%	90%	95%	80%	90%	95%	
1	2280	2340	2370	2400	2580	2760	2910	3,000
2	2280	2340	2370	2400	2580	2760	2910	3,000
3	1830	1890	1890	1920	2070	2190	2310	2,400
4	1830	1890	1890	1920	2070	2190	2310	2,400
5	1560	1590	1590	1620	1770	1890	1980	2,400
6	1560	1590	1590	1620	1770	1890	1980	2,400
7	1560	1590	1590	1620	1770	1890	1980	2,400
8	2070	2130	2160	2190	2340	2490	2640	2,610
9	2070	2130	2160	2190	2340	2490	2640	2,610
10	1770	1830	1860	1860	1980	2130	2250	2,220
11	1980	2040	2070	2100	2250	2400	2550	2,580
12	1980	2040	2070	2100	2250	2400	2550	2,580
13	1680	1740	1770	1770	1920	2040	2160	2,250
14	2790	2910	2910	2940	3180	3390	3570	3,420
15	2790	2910	2910	2940	3180	3390	3570	3,420
16	1830	1890	1920	1920	2070	2220	2340	2,400
17	1830	1890	1920	1920	2070	2220	2340	2,400
18	1590	1620	1650	1650	1800	1920	2040	2,220
Mean	**1960**	**2020**	**2038**	**2060**	**2221**	**2370**	**2501**	**2595**

In reality, the methods can be compared in terms of the number of households to sample, but it is preferable to check whether the yield of children under 5 years of age is adequate, because this is the ultimate aim of the predictions. Therefore, as before, the results in Table 12 are generated by multiplying the results in Table 11 by the actual kappa value (κ_0), the value realized through survey work. The results in last column (Stukel-Deitchler method) are obtained from the last column of Table 2.

Table 12. Expected yield of children under 5 years of age using the four methods (Magnani, Poisson, Kappa Prediction, Stukel-Deitchler)

Survey Number	Required Sample Size of Children (*n*)	Magnani	Poisson			Kappa Prediction			Stukel-Deitchler
		50%	80%	90%	95%	80%	90%	95%	
1	1694	2316.5	2377.4	2407.9	2438.4	2621.3	2804.2	2956.6	3,048
2	1694	2000.3	2053.0	2079.3	2105.6	2263.5	2421.4	2553.0	2,632
3	1557	2213.5	2286.1	2286.1	2322.4	2503.8	2649.0	2794.1	2,903
4	1557	2103.7	2172.7	2172.7	2207.2	2379.6	2517.6	2655.5	2,759
5	1686	2717.2	2769.5	2769.5	2821.7	3082.0	3291.0	3447.7	4,179
6	1686	2029.1	2068.2	2068.2	2107.2	2303.9	2460.1	2577.3	3,124
7	1686	1996.3	2034.7	2034.7	2073.1	2265.6	2419.2	2534.4	3,072
8	1227	1491.8	1535.1	1556.7	1578.3	1686.4	1794.5	1902.6	1,881
9	1227	1510.1	1553.8	1575.7	1597.6	1707.0	1816.5	1925.9	1,904
10	980	1161.7	1201.0	1220.7	1220.7	1299.5	1397.9	1476.7	1,457
11	1454	1521.8	1568.0	1591.0	1614.1	1729.4	1844.7	1959.9	1,983
12	1454	1442.8	1486.5	1508.4	1530.2	1639.5	1748.8	1858.1	1,880
13	1434	1485.1	1538.2	1564.7	1564.7	1697.3	1803.4	1909.4	1,989
14	1380	1086.6	1133.4	1133.4	1145.1	1238.3	1320.0	1390.1	1,332
15	1380	1665.0	1736.6	1736.6	1754.5	1896.9	2022.1	2129.5	2,041
16	1442	1641.7	1695.5	1722.4	1722.4	1857.1	1991.6	2099.3	2,153
17	1442	1292.4	1334.8	1356.0	1356.0	1461.5	1567.4	1652.1	1,695
18	1627	1979.2	2016.5	2053.9	2053.9	2240.4	2389.7	2539.1	2,762
Mean	**1478.2**	**1759**	**1809**	**1824**	**1845**	**1993**	**2125**	**2242**	**2,377**

In comparing all methods, it is interesting to note that the projections of the expected yield of children increase in value moving from the leftmost method (Magnani) to the penultimate rightmost method (Kappa Prediction at 95% confidence level). The last method (Stukel-Deitchler) behaves somewhat like the Kappa Prediction method but oscillates above and below the 95% confidence level.

As before, underestimation in the yield of children required is indicated in red. It is evident with these data that a minimum 95% confidence level is needed under the Kappa Prediction method to avoid any underestimation in the number of households to sample. However, this is driven by a single survey, Survey 14, where the estimate of kappa was off by a singularly large 27% and is ranked number 1 in terms of the ratio of the estimated kappa to the actual kappa in Table 10. Removing that survey from the analysis, it is possible to use the Kappa Prediction method at the 80% confidence level to yield results having no underestimation. Furthermore, if we remove Survey 17, which also had a very large (12%) overestimation in Kappa and is ranked number 2 in terms of the ratio of the estimated kappa to the actual kappa in Table 10, it is possible to use the Poisson method at the 80% confidence level to yield results having no underestimation, and, in doing so, saving considerable cost in terms of households to sample. More generally, if kappa can be estimated perfectly (rarely the case), then the Magnani method is the preferred approach (see Table 4). Otherwise, it may be prudent to build in a small cushion of extra sample size by using the Poisson method at 80% confidence level – at least for the 18 surveys at hand.

6. Conclusions and Recommendations

This paper tackles the problem of appropriately projecting the number of households to sample in order to achieve a required target sample size of eligible children within the sampled households. There are a number of methods that can be used to inflate the required number of children (once established) to the commensurate number of households. This paper considers four such inflation methods: Magnani, Poisson, Kappa Prediction, and Stukel-Deitchler. However, all methods rely on knowledge of two input parameters: lambda (the average number of eligible children per household that respond) and gamma (the average household response rate among sampled households), the product of which is called kappa. Because these input parameters are unknown prior to survey implementation, they must be estimated. However, the input parameters are often not very accurately estimated in practice, despite the fact that the ability to accurately project the number of households to sample to achieve the required target number of children is highly sensitive to accurate estimation of these input parameters.

This paper begins by simulating a scenario of perfect prediction of these input parameters, lambda and gamma (the product of which is kappa). Under this scenario, the Magnani method provides the most accurate projection of the number of households to sample, assuming that it is always preferable to modestly overestimate the number of households required rather than modestly underestimate the number of households required. Table 4 shows that, using the results of 18 USAID/FFP baseline surveys, the Magnani method in fact only very modestly overestimates the projected number of households to sample (and by extension the required sample size of children) in each case—and therefore is the most cost-effective in terms of sample size savings. Therefore, the Magnani method is the preferred method to use, provided that the input parameters can be perfectly estimated prior to survey implementation.

However, in most cases, the input parameters are inaccurately estimated, and sometimes very inaccurately estimated. This is a critical issue and must be addressed. For the Magnani method, when kappa is underestimated prior to survey work, the inflation methods will lead to overestimation in the number of households to sample (and hence the number of children on which data are collected). Conversely, when kappa is overestimated prior to survey work, the methods will lead to underestimation in the number of households to sample. Neither scenario is desirable, although modest overestimation in the number of households to sample (and hence, children) is preferable to underestimation. This paper uses the fact that the input parameters from the 18 USAID/FFP baseline surveys have been inaccurately estimated to simulate the behavior of the four proposed methods under this scenario. Table 12 demonstrates that, after removing two outlier surveys from the analysis, the Poisson method at the 80% confidence level builds a small cushion of excess sample (compared to the Magnani method) and so the overestimation in the number of households to sample is modest. Therefore, under the scenario where the input parameters are inaccurately estimated, the Poisson method at the 80% confidence level is the preferred method to use for this particular set of 18 USAID/FFP baseline surveys. Both remaining methods (Kappa Prediction and Stukel-Deitchler) inflate the sample size excessively compared to the Magnani and Poisson methods for the 18 surveys in question.

However, to extend the recommended use of these methods beyond the 18 USAID/FFP baseline surveys more generally is complex, because it is unknown in advance of obtaining survey results what will be the extent and size of under- or overestimation in the kappa parameters. Therefore, in a more general setting beyond the 18 surveys at hand, a more conservative approach should be adopted; that is, in such cases, it is recommended that the Poisson method at the 95% confidence level be adopted to allow for a greater sample size of households (over the Poisson method at the 80% confidence level). This will provide a greater cushion to protect against not achieving the target number of eligible children required by the survey.