

# Correcting Survey Non-Response With Census Data

---

Johannes G. Hoogeveen<sup>1</sup> and Youdi Schipper<sup>2</sup>

## Summary

*Household size related non-response occurs because the enumerator fails to find someone at home, or because information on household members is not captured. Using census and survey data from Uganda we show how such non-response leads to substantial bias in the survey distribution of household size and how, with the use of commonly available census information, the non-response bias can be corrected.*

## Keywords

*Poverty and inequality measurement.*

## Résumé

*La non-réponse liée à la taille du ménage se produit quand l'agent recenseur ne trouve personne à la maison ou quand l'information sur les membres du ménage n'a pas été saisie. En utilisant des données de recensement et d'enquête de l'Ouganda, nous montrons comment ces non-réponses occasionnent des biais importants dans la distribution de la taille des ménages issue des enquêtes et comment corriger ces biais en utilisant des informations du recensement communément disponibles.*

## Mots clés

*Pauvreté et mesure d'inégalités*

## 1. Introduction

This paper considers how to correct for non response in a sample survey when non-response is related to household size. Household size related non-response may occur when an enumerator fails to find any respondent at home. The probability of

---

1: Johannes Hoogeveen is with the World Bank, C. O. World Bank, P.O.Box 2054, Dar es Salaam, Tanzania, jhoogeveen@worldbank.org.

2: Youdi Schipper works with the Vrije Universiteit Amsterdam, FEWEB-4A29 De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, yschipper@feweb.vu.nl. We are grateful to the Uganda Bureau of Statistics for their help with the provision of survey and census data. We also very much appreciate useful help and suggestions from Johan Mistiaen. The findings, interpretations and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the view of the World Bank, its Executive Directors, or the countries they represent.

3: It would appear that the probability of not finding the household head at home is much more equally distributed over household size classes. However, if the head of household is not at home, typically another household member will respond.

finding no one at home in a one person household is larger than in a multiple person household.<sup>3</sup> Small households are therefore likely to be underrepresented in the survey. A second type of household size related non-response results from the non-recording of individual household members. In surveys where information has to be collected for each individual in the household (as is typical for living standards type of surveys) not all individuals may be captured. This is more likely to happen in very large households. Both types of household size related non-response lead to an under-representation of both small and very large households and an over-representation of medium sized households.

If non-response is correlated with household size and if non-responding households are replaced in the sample without taking this correlation into account, substitute households are likely to differ systematically from the non-respondents they replace (Kish, 1965, Lessler and Kalsbeek, 1992). As a result, survey estimates may be biased. This paper presents an ex-post approach to correcting for household size related non-response by using an external source of data, the population census.

This paper follows a tradition in which adjustments for non-response are made by identifying (re)weighting factors for every household in the survey. Various methods for determining these factors have been suggested. One proposal has been to infer the weights using partial information that may exist on non-respondents including, for instance, the number of attempts required to obtain a response (Politz and Simmons, 1949). An alternative method infers these weights from the distribution of non-respondents across certain identifiable subgroups of the sample, called "adjustment cells" (Thomsen, 1973). External sources of data, such as a population census, have also been proposed to determine the 'true' number of units in the various subgroups of the population (Hansen et al., 1953). Our contribution in this paper is to clarify how non-response may be related to household size and to show how external sources of data can be used to correct for such non-response. We also show how these external data can be used to assess whether adjustment reduces survey bias in dimensions beyond the one (household size) for which the correction is carried out.

We address these issues using the 1992 Uganda Integrated Household Survey (IHS) and the 1991 population census. We show how survey non-response varies with household size using the "correct" census distribution of household size (Section 2) and adjust survey weights accordingly (Section 3). We then show that using the new weights, the comparability between census and survey improves considerably. Section 4 investigates the bias that household size related non-response may create for commonly used welfare indicators, explores the importance of income re-

lated non-response, and shows that the problem is not confined to Uganda. Concluding remarks, finally, present a case to experiment with different visiting strategies to altogether avoid household size related non-response.

### 2. Survey Non-response and Household Size

Survey non-response is likely to be highly correlated with household size as the probability of finding someone at home is likely to be higher for large than for small households. With a fixed number of attempts to visit a household small households are disproportionately likely to not be included in a survey. Formally, let the probability of a respondent  $i$  being at home at the time an enumerator calls be represented by  $p_i$  and assume that the presence of household members at home is independent. The probability of non-response  $p_s^n$  for a household with  $s$  number of respondents can then be expressed as:

$$p_s^n = (1 - p_i)^s$$

Expression (1) is non-linear and the probability of non-response decreases exponentially with the number of respondents. Consider a probability of a respondent being at home of, say, 0.7. This probability is probably an over-estimation, considering that it reflects the probability that an adult household member who is able to respond is at home at a time when enumerators carry out their assignment. At this relatively high probability of being at home, the probability of non-response drops to below 3 percent for households of size 3 and larger. But for small households (of course depending on the probability of finding a respondent at home) the probability of non-response is still non-negligible. In our illustration but with only one attempt to visit a household the probability of non-response for a one-respondent household is 30 percent and for a household with two respondents 9 percent. If the number of visits is increased these fractions will diminish, but a sample survey that does not have a visiting strategy which increases the likelihood of finding a respondent at home to close to unity and whose replacement procedure is independent of household size is likely to under-sample small households and over-sample larger households. Without knowing the probability  $p_i$ , the final dataset cannot be corrected for this bias.

Another reason leading to the incorrect representation of household size in surveys is that household members may not be recorded; this may be due to memory lapse or enumerator error. Let the probability of non-recording be  $r_i$  for each household member. Given that an enumerator is interviewing one household member, the prob-

ability that at least one household member is not recorded in a household of size  $s$  can be expressed as one minus the probability that all remaining household members are recorded:

$$p_s^r = (1 - r_i)^{s-1}$$

This probability increases with household size at a decelerating pace and approaches 1 for very large households. Even a low probability of non-recording of say 0.01 will lead to an almost 5 percent probability that household size is underestimated by at least one member for a household whose (true) size is 6 and of over 10 percent for a household whose (true) size is 12.

The probability of non-response and the probability of under-recording in a sample survey lead to the tails of the distribution of household size being thinner than they should be. Since both types of measurement problems have potentially offsetting consequences for the mean or the median, the thinning of the tails in the sample survey could easily go undetected without considering the distribution of households over size classes. By its very nature a census focuses on the correct administration of household size. Hence much attention is devoted to training of enumerators to ensure that this aspect of the census is 'right'. Also, census enumeration takes little time so that inclusion of an extra individual has small marginal cost for the enumerator and respondent. In contrast, household survey questionnaires typically take much more time and are less focused on obtaining correct household size. Provided, then, that a census is not subject to non-response problems the 'thinning' of the tails of the household size distribution in survey data can be identified by comparing both distributions.

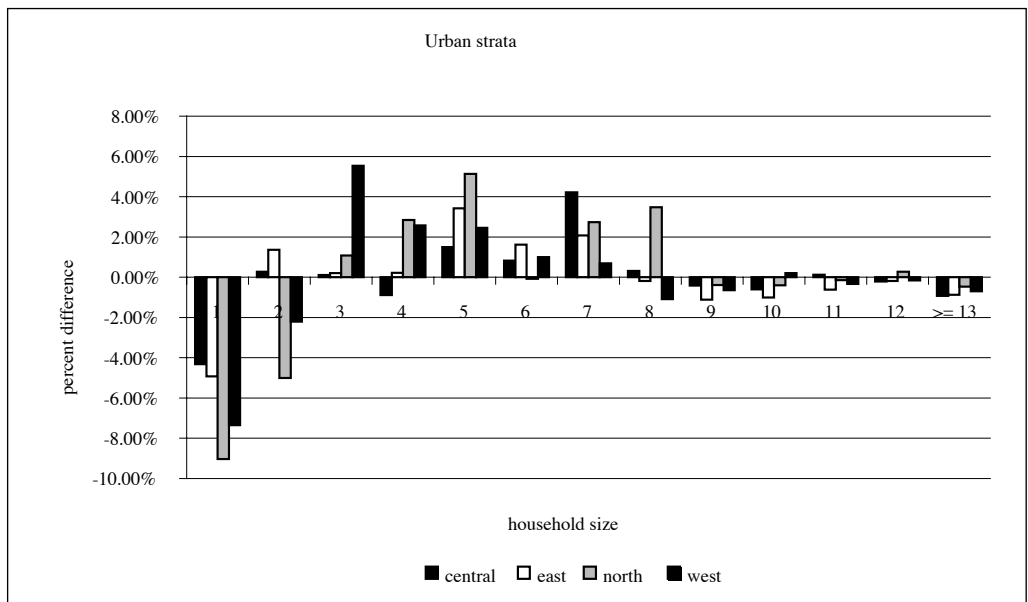
How important are these effects in practice? In the second week of January 1991 a population census was held in Uganda, whereas from January to December 1992 the Integrated Household Survey was implemented. Both the survey and the census recorded household size by collecting information on each member of the household. The definition for a household is identical between the survey and the census.<sup>4</sup> The survey was representative in 8 strata (rural and urban areas in respectively central, east, north and west Uganda), so that measures of household size should be (statistically) identical between the census and the survey.

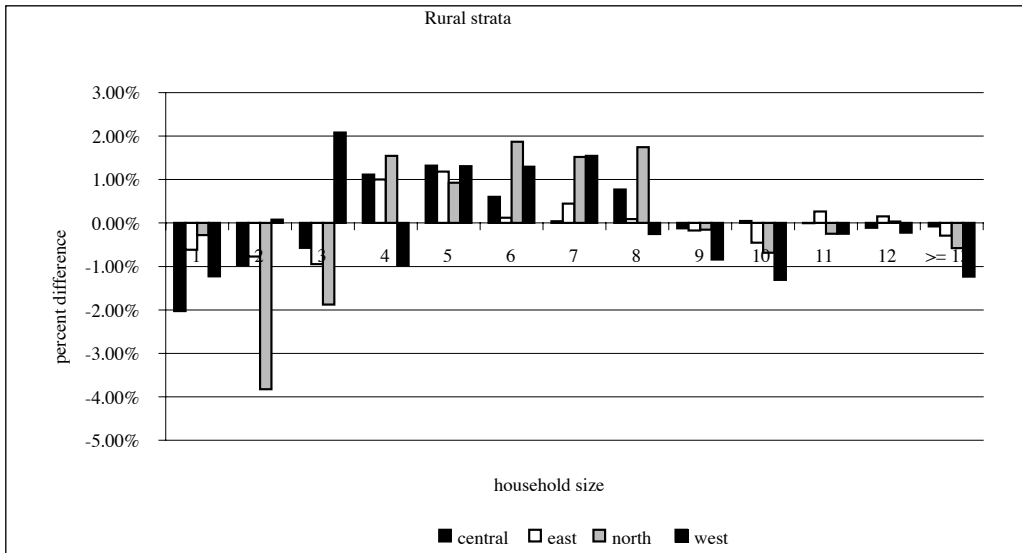
---

1: A household is defined as a group of persons who live, cook and eat together or a single person who lives alone and eats independently.

Figure 1 explores the existence of a bias in greater detail. It plots for each stratum and for different household size-categories the difference in the proportion of households reported by the census and the survey. In total 13 categories are distinguished: households of size 1 up to 12 (reflecting 98 percent of the total population) and a rest category comprising all households with 13 or more members. The figure shows a clear pattern. Households comprising 1 or 2 members (in rural areas 1, 2 or 3 members) are under-represented in the survey; households of size 4-8 are over-represented. Large households comprising 9 or more members are under-represented again. Not only is the pattern striking, the differences can be sizeable especially in urban areas where, depending on the stratum, the fraction of one-person households in the survey is four to more than eight percentage points less than that in the census. In rural areas the differences are less pronounced, yet the pattern is comparable to that in urban areas.

**Figure 1: Difference in proportion of households of size x reported in census and sample survey (a negative number indicates under-representation in the survey).**





### 3. Adjusting for Survey Non-response

Having established the existence of the problem, the question is “can we control for it?” In the presence of a census, this can be achieved through a re-weighting procedure which ensures that the frequency distribution among mutually exclusive and exhaustive categories in the survey correspond precisely to the frequency distribution among those same categories in the census. Formally, let the total population in the census be  $N$  and let there be  $N_s$  people in the census living in households of size  $s$ . Corresponding variables for the survey are indicated by a lower case. If the inflation factor for a household  $i$  of size  $s$  is  $w_{is}$ , then, according to a survey of  $h$  households,

the total population living in households of size  $s$  is given by  $n = \sum_{i=1}^h s_i w_i$  and the total population is:  $n_s = s \sum_{i=1}^{h_s} w_{is}$ . Denote the fraction of the population living in a size  $s$

household as  $f_s$ , which corresponds to the fraction  $F_s$  that can be obtained from the census as  $N_s / N$ . If, as is the case in Uganda, the total population according to the survey ( $n$ ) corresponds to the total population according to the census ( $N$ ), while the distribution of the population over different household size-categories does not correspond to that of the census, then this can be corrected by multiplying  $w_{is}$  by  $N_s / n_s$ .

Though the procedure to correct for a bias due to household size related non-response is straightforward, in practice there may be reluctance to rely on re-weight-

ing because adjusting a survey in one dimension may make it less comparable to a census in others. However, if the adjustment is to deal with size bias due to non-response, then adjusting the survey weights should not only lead to improvements in comparability in household size, but in other dimensions as well. By considering whether re-weighting results in improvements – or at least not in deterioration – in the comparability with other variables, we have a test for the validity of the exercise. Clearly this test can only be applied if the census and survey reflect the same time period (which is the case in Uganda), and if variables are identically defined.

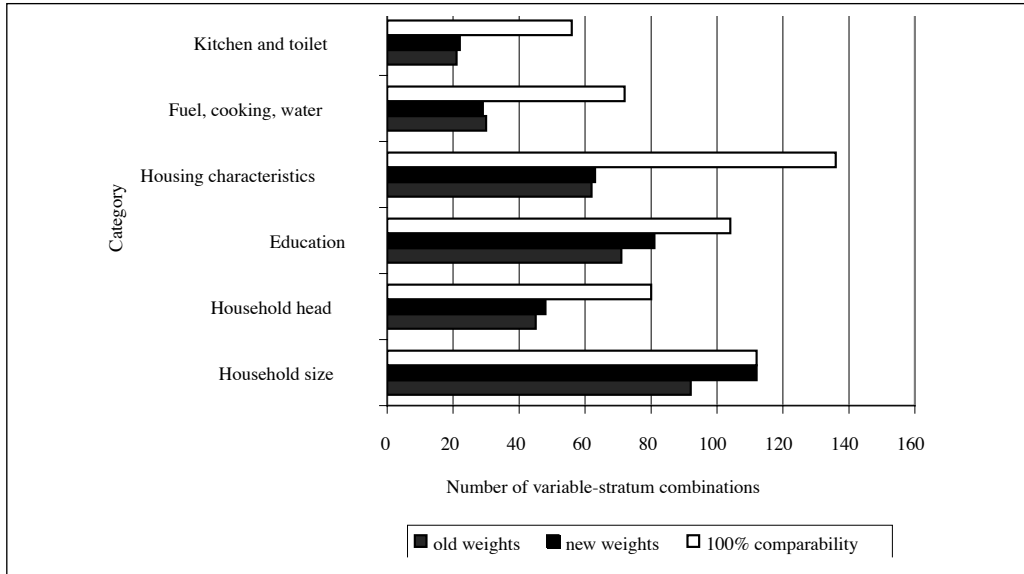
After comparing enumerator instructions and the way questions were phrased and coded between the census and the survey, 70 potentially identical variables were identified.<sup>5</sup> As the survey is representative at the stratum level (there are 8 strata), there are 560 possible variable-stratum combinations for which the census mean should lie within the 95% confidence interval of the survey. Having sorted the 70 variables into six variable categories, we present the number of ‘matching’ variable-stratum combinations by variable category in Figure 2. The figure also indicates the maximum number of matches attainable (‘100% comparability’). By definition, after re-weighting all variables relating to household size pass the comparison test and comparability increases from 321 variable-stratum combinations to 355 (+10.6%). Comparability for the non-household size related variables increases from 229 to 243 variable-stratum combinations (+ 6.1%).<sup>6</sup> We therefore conclude that household size related non-response did affect the IHS and that the re-weighting procedure resulted in an improved set of household weights.

---

5: Even when variable definitions, enumerator instructions and coding are identical, different responses may arise, due to different enumerator training procedures, or because of minor differences in response codes.

6: Only with respect to two variables, marital status (head of household never married) and fuel use (household cooks using paraffin) does comparability deteriorate by one variable-stratum combination.

**Figure 2: Number of variable-stratum combinations for which census mean falls within survey mean's 95% confidence interval**



#### 4. Discussion

Our analysis has shown how household size related non-response may lead to an unrepresentative sample and, consequently, biased estimates. Table 1 provides insight into the importance of re-weighting for poverty incidence, per capita consumption and the Gini coefficient.

**Table 1: Comparison of consequences of reweighting on various indicators of household welfare**

Domain	Poverty Incidence		Per capita consumption		Gini Coefficient	
	IHS, official	IHS, re-weighted	IHS, official	IHS, re-weighted	IHS, official	IHS, re-weighted
Urban	27.8 (2.4)	27.8 (2.4)	33158 (2063)	32534 (1699)	0.395 (0.03)	0.383 (0.03)
Central rural	54.3 (2.2)	54.1 (2.2)	18046 (638)	18131 (629)	0.329 (0.01)	0.330 (0.01)
East rural	60.6 (2.3)	60.6 (2.3)	15427 (480)	15460 (486)	0.321 (0.01)	0.322 (0.01)
North rural	73.0 (2.9)	72.3 (2.9)	13663 (632)	13899 (636)	0.330 (0.02)	0.331 (0.01)
West rural	54.3 (2.4)	55.0 (2.6)	16368 (500)	16256 (537)	0.309 (0.01)	0.311 (0.01)

**Notes:**The columns IHS official presents welfare estimates as released by the Uganda Bureau of Statistics. IHS re-weighted adjusts the IHS sampling weights for household size related non-response. Standard errors are in parentheses and are corrected for survey design.

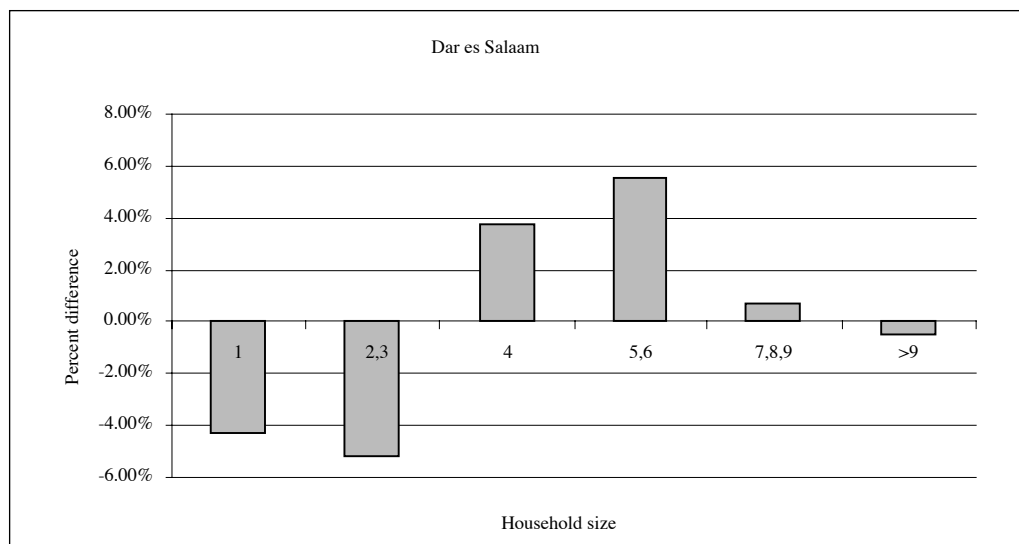
The table shows how re-weighting the IHS to adjust for household non-response has little effect on the various welfare estimates. This is encouraging in that it implies that official poverty estimates do not need to be revised. This absence of an impact of re-weighting on the welfare indicators can be traced to two aspects: (i) the fraction of poor one and two person households is small; and (ii) even after re-weighting, members from small households make up only between 8% and 9% of the total population. It should not be inferred from this lack of impact that re-weighting is superfluous. Whether this is the case depends on the research question at hand. For instance, if the interest were in the fraction of non-poor living in small households then re-weighting makes a significant difference (at the 95% level of confidence) as it increases the fraction from 39.3% to 45.1%.

Non-response is not related to household size alone. A distinct concern may be that survey participation varies with household income. Very poor and very rich households, for instance, may be less inclined to answer because of the high opportunity costs of their time. Household size related non-response may also be correlated with income. This could occur, for instance, when non-responding households are more likely to be non-poor because they are employed (and hence not at home when

the enumerator visits). In the latter case, correcting for non-response with the procedure outlined previously will not correct this bias. Mistiaen and Ravallion (2003) show how, in the presence of information about non-response rates, income related non-response may be corrected for. They present an application for the USA, showing a considerable under-representation of wealthy households. Applying their approach to the Ugandan survey we obtain an inverted-U shaped compliance-expenditure pattern with people in middle quintile groups more likely to comply than either the richest or the poorest. The difference in compliance rates is only marginal. After correcting for wealth related non-compliance the largest divergence we find for the poorest quintile an estimated true population proportion of 0.2097 (rather than 0.20); for the wealthiest quintile it is 0.1986. On the basis of this information, we conclude that household size related non-response is a more pressing issue than income related non-response.

So far we illustrated the importance of household size related non-response with data from Uganda. However, the issue is not unique to Uganda. In Tanzania, for instance, a pattern comparable to that presented in Figures 1 and 2 was detected when comparing the distribution of household sizes in the 2000/01 Household Budget Survey with that of the 2002 Population and Housing Census. There is an under-representation of small and very large households and an over-representation of medium sized households. Figure 3 illustrates the Tanzania pattern for its capital city, Dar es Salaam. We therefore conclude that many more surveys are likely to be affected by this easily remedied problem.

**Figure 3: Difference in proportion of households of size x reported in the Tanzanian 2002 census and the 2000/01 Household Budget Survey (a negative number indicates under-representation in the survey).**



**Source:** Adjusted from Kilama and Lindeboom (forthcoming)

### 5. Concluding Remarks

Our analysis has shown how household size related non-response may lead to an unrepresentative sample and, consequently, biased estimates. A straightforward way to deal with the bias has been presented in which the frequency distribution of household size categories in the survey is made to correspond precisely to the frequency distribution among those same categories in the census.

The problem of household size related non-response can largely be avoided by re-considering the visiting strategy of enumerators. If enumerators make various attempts to enumerate a household and (re)-visit on different days of the week and at different times during the day, the probability of non-response by small households is likely to decline. Little can be said, a priori, which visiting strategy would be optimal. Yet a survey that would experiment with visiting strategies that vary across strata and which checks, ex post, the stratum level frequency distributions of household size in the survey and with those from the census would provide valuable information that would help avoid this kind of bias in the future. Identifying a visiting strate-

gy that would avoid household size related non-response is critical, especially as the ex post procedure sketched in this paper is only feasible when the survey and census are contemporaneous. As censuses are only implemented once every ten years, most surveys will not benefit from the procedure. Improving the household visiting strategy and thus reducing household size related non-response therefore remains an important objective to pursue.

## References

Hansen, H.M., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. John Wiley & Sons, New York

Kilama, B. and Lindeboom, W. *Where are the Poor in Tanzania* (forthcoming).

Kish, L. (1965). *Survey sampling*. John Wiley & Sons, New York.

Lessler, J. T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. John Wiley & Sons, New York.

Mistiaen, J. and Ravallion, M. (2003). *Survey Compliance and the Distribution of Income*. World Bank: Policy Research Working Paper no. 2956.

Politz, A.N. and Simmons, W.R. (1949). An Attempt to Get 'not-at-homes' Into the Sample Without Call-backs, *Journal of the American Statistical Association*, 44, 9-31.

Thomsen, I. (1973). A Note on the Efficiency of Weighting Subclass Means to Reduce Effects of Non-response When Analyzing Survey Data, *Statistik Tidskrift*, 4, 278-283.