Mismeasured Household Size and Its Implications for the Identification of Economies of Scale

By Timothy J. Halliday,*

Department of Economics and John A. Burns School of Medicine,

University of Hawaii at Manoa

Abstract

We consider the possibility that household demographic variables are measured with error. Such errors will arise because income and consumption surveys measure the household's structure at a point-in-time, whereas the demographic composition of the household is constantly evolving over the survey period. We construct and estimate sharp bounds which suggest that the degree of these measurement errors is non-trivial. We also provide evidence that these errors may have important ramifications for recent work on the identification of economies of scale within households.

---

1   Introduction

In this paper, we consider the possibility that household demographic variables are measured with error. Such measurement errors may arise because the bulk of consumption and income surveys in the world ask households to report their demographic structures at a point-in-time. These data are then typically used by researchers to proxy for the household's structure over the duration of the survey period. However, if the household undergoes any number of demographic changes such as migration, fertility, mortality, marriage and/or divorce during the survey period then information reported at the time of data collection may deviate substantially from the household's average demographic structure during the period.

To provide readers with some notion of how fluid household structures may induce measurement errors in demographic variables, we use data from the BASIS Panel in El Salvador and the Panel Study of Income Dynamics (PSID) in the United States to calculate the change in household size across survey years.[1] Figures 1 and 2 show changes in El Salvador and the United States, respectively. The figures suggest that household structures are fluid. In El Salvador, fewer than 50% of households experienced no change in household size across survey years. In the United States, this figure rises to about 83%. We therefore expect measurement errors in demographic variables to be present for over 50% of households in the Salvadoran data and about

---

[1] The BASIS panel covers the years 1997, 1999 and 2001. The PSID data covers the years 1990, 1991, 1992 and 1993.

17% of households in the American data.

Unfortunately, surveys do not collect information on the household's demographic structure at all points-in-time during the survey period. Consequently, researchers are unable to accurately calculate the household's average size over the survey period. In this paper, we try to gain some insight into the extent of this problem. Following Manski (2003), we impose weak assumptions on the demographic processes which took place within the household during the survey period and then use these assumptions to derive bounds on expectations of average household size during the survey period. The derived bounds are sharp which is to say that they cannot be improved without stronger assumptions. We calculate these bounds using the BASIS panel from El Salvador, a country in which a large amount of trans-national migration takes place.

Our calculations show that these bounds are often wide. For households with two migrants residing abroad, these bounds indicate that average household size may fluctuate by more than one household member. When we consider households with three or more migrants, these calculations indicate that average household size may fluctuate by more than two household members. About 15% of the households in our data report having at least two migrants residing abroad and, thus, these bounds are very wide for a non-trivial portion of our data. One important caveat is that while large bounds do suggest that measurement errors in household demographics are important, they are necessary, but not sufficient, for measurement errors to be present.

These findings may have ramifications for recent work by Deaton and Paxson (1998) (DP) who develop a test for economies of scale within the household. Their test hinges on the observation that intra-household public goods become cheaper as households become larger; in effect, larger households are richer conditional on *per capita* expenditures. Accordingly, if we condition on

per capita expenditures, we should observe that the consumption of goods with sufficiently high income effects (such as food in developing countries) increases with household size. However, DP present comprehensive evidence that directly contradicts this prediction. Moreover, the contradiction is the strongest for the poorest countries in which we would expect the income elasticity of food to be the highest.

Measurement errors in household size may be able to explain at least part of the DP paradox. The reason is that their tests include both *per capita* consumption expenditures and household size on the right-hand size of the regression equation. We show that this will induce a bias in the estimate of the economies of scale coefficient which will be an increasing function of the coefficient on *per capita* expenditures. Because Engel's Law implies that this coefficient is highly negative, we should expect to see a severe negative bias in the estimate of the economies of scale coefficient.

The balance of this paper is organized as follows. In Section 2, we formally state the problem. In the next two sections, we show how weak and strong assumptions allow us to derive bounds on expectations of average household size during the survey period. In Section 5, we describe the data. In the next two sections, we discuss estimation and inference issues and then present our results. In Section 8, we discuss the ramifications of mismeasured household size for the identification of economies of scale. Section 9 concludes.

2   The Problem

We assume that the household's decision process unfolds in continuous time. We let $N(s)$ denote the household's size at time s. We remain agnostic about the household's underlying decision

process and we assume that $N(s) \geq 1$ for all s.

Unfortunately, survey instruments only collect data at discrete intervals such as once every year or every two years. As a consequence, researchers do not observe the quantity $N(s)$ for all s in any given time interval. This forces researchers to summarize $N(s)$ over discrete time intervals such as $[t-1, t]$ or $[t, t+1]$.

To help fix ideas, we define the object

$$N_t^* \equiv E[N(s)|s \in [t-1, t]] \tag{1}$$

which denotes the average of the household's size over the interval $[t-1, t]$. Note that this expectation is taken across time for a given household and, thus, there will be a distribution of N_t^* across households. Because most household surveys do not permit the precise measurement of N_t^*, researchers typically proxy for the household's size over the survey period with $N_t = N(t)$ where t is the time that the survey was administered. Unfortunately, this can be problematic as the household's structure often changes during the survey period as a consequence of birth, death, migration, marriage and/or divorce. In other words, N_t^* and N_t may substantially deviate from each other. When this occurs, household size will be measured with error which can be written as

$$\epsilon_t = N_t^* - N_t. \tag{2}$$

If the household's demographic structure is constant over the time interval so that $N(s) = N_t$ for all $s \in [t-1, t]$, then there will be no measurement error and, $N_t^* = N_t$. Otherwise, errors will be present. This could be particularly problematic for developing countries such as El Salvador

where there is a tremendous amount of fluctuations in household size as indicated by Figure 1.

3 A Bounds Analysis with Weak Assumptions

We now show how we can use information from our surveys to construct bounds on expectations of N_t^*. If there are measurement errors in demographic variables then these bounds will be wide. However, the converse is not true. Thus, wide bounds suggest, but do not imply, that measurement errors are problematic.

We now introduce some notation. We let M_t denote the number of migrants in the household at the time of the survey, t. We define a migrant to be a household member residing outside of the household's dwelling. It is important to note that N_t only includes home dwellers and not migrants. We let B_t denote the number of births that took place in the household during the survey year. Finally, we let D_t denote the number of deaths which took place during the survey year. Throughout this section, we do not address marriage or divorce.

The following identity holds:

$$N_t = N_{t-1} - \Delta M_t + B_t - D_t. \tag{3}$$

We emphasize that the quantity ΔM_t is *net* migration. Based on this simple identity, we propose some sensible assumptions which will allow us to construct bounds on expectations of $N(s)$ for $s \in [t-1, t]$.

Suppose that the only demographic change that takes place in the household over the survey period is migration. Then, we will have that $N_{t-1} = N_t + \Delta M_t$. If $\Delta M_t > 0$, then this implies

that

$$N_t < N_{t-1} = N_t + \Delta M_t. \tag{4}$$

Then we may reasonably assume that $N(s)$ was in the interval $[N_t, N_t + \Delta M_t]$ for all $s \in [t-1, t]$.

We use this logic to make three assumptions on the process for $N(s)$:

$$N(s) \in [N_t - B_t, N_t + D_t] \text{ for } \Delta M_t = 0 \text{ and } s \in [t-1, t], \tag{W1}$$

$$N(s) \in [N_t - B_t, N_t + D_t + j] \text{ for } \Delta M_t = j > 0 \text{ and } s \in [t-1, t] \tag{W2}$$

and

$$N(s) \in [\max\{N_t - B_t + j, 1\}, N_t + D_t] \text{ for } \Delta M_t = j < 0 \text{ and } s \in [t-1, t]. \tag{W3}$$

The lower bound in A3 results from the assumption in Section 2 that the household size is always positive at any point-in-time. This is the weakest set of assumptions that we employ. In the next section, we show how a progressively stronger set of assumptions can be used to derive tighter bounds and, ultimately, achieve point-identification.

It is important to emphasize that these conditions are assumptions and are not simply implied by the identity in equation (3). To better understand this, we consider a hypothetical scenario in which the household size was five at the end of the survey period and net migration out of the household was two during the survey period. For the sake of simplicity, we assume that no births or deaths took place during the survey period. In this scenario, Assumption W2 implies that $N(s)$ will lie in the interval $[5, 7]$ for all $s \in [t-1, t]$. However, in the absence of Assumption W2, this need not be the case. Because ΔM_t is net migration over a time interval, it may mask movements

in the household's demographic structure which occur within time intervals. Going back to our example, it could be that, just after time $t-1$, four members migrate out of the household and just prior to the end of the survey, at time t, two of these same members subsequently return to the household. In this hypothetical case, net migration out of the household would still be two. However, $N(s)$ would be in the interval $[3, 7]$ not $[5, 7]$. Assumptions W1 through W3 rule out these types of scenarios.[2]

These assumptions can easily be used to construct bounds on the conditional expectation of average household size: $E[N_t^*|W_t]$ where $W_t \equiv (N_t, M_t, D_t, B_t)$. We choose to bound $E[N_t^*|W_t]$ because many surveys contain information on W_t whereas fewer surveys contain information on ΔM_t. Accordingly, the calculations in this paper shed light on measurement errors in a broader set of surveys than if we were to also condition on ΔM_t. Finally, it is important to note that the expectation, $E[N_t^*|W_t]$, is taken over households whereas the expectation, N_t^*, is taken over time for a given household.

We now derive the bounds in a series of steps. First, we note that the assumptions on $N(s)$ imply the following bounds on N_t^*:

$$N_t^* \in [N_t - B_t, N_t + D_t] \text{ for } \Delta M_t = 0 \tag{5}$$

$$N_t^* \in [N_t - B_t, N_t + D_t + j] \text{ for } \Delta M_t = j > 0 \tag{6}$$

---

[2]While we concede that these assumptions may be unrealistic in certain circumstances, they are still far weaker than the assumption that the household's demographic structure was constant over the survey period which is an assumption is employed in the vast majority of studies. Thus, it is impossible to take exception to assumptions W1 through W3 without taking exception with the implicit assumptions in much of the literature.

and

$$N_t^* \in [\max\{N_t - B_t + j, 1\}, N_t + D_t] \text{ for } \Delta M_t = j < 0. \tag{7}$$

Second, we note that, by the Law of Iterated Expectations, we can write

$$E[N_t^*|W_t] = \sum_j E[N_t^*|\Delta M_t = j, W_t] P(\Delta M_t = j|W_t). \tag{8}$$

Third, conditions (5), (6) and (7) imply that

$$N_t - B_t \leq E[N_t^*|\Delta M_t = j, W_t] \leq N_t + D_t + j \text{ for } j > 0, \tag{9}$$

$$N_t - B_t \leq E[N_t^*|\Delta M_t = j, W_t] \leq N_t + D_t \text{ for } j = 0$$

and

$$\max\{N_t - B_t + j, 1\} \leq E[N_t^*|\Delta M_t = j, W_t] \leq N_t + D_t \text{ for } j < 0. \tag{10}$$

These bounds together with equation (8) give us Proposition 1.

Proposition 1 *If the process of $N(s)$ satisfies assumption W1 through W3 then we will have that*

$$L(W_t) \leq E[N_t^*|W_t] \leq U(W_t)$$

where

$$U(W_t) \equiv N_t + D_t + \sum_{j>0} j * P(\Delta M_t = j|W_t)$$

9

and

$$L(W_t) \equiv N_t - B_t + \sum_{j<0} \max\{j, 1 - N_t + B_t\} * P(\Delta M_t = j|W_t)$$

An important question is whether or not we can improve upon the bounds in Proposition 1 while only maintaining assumptions W1 through W3. In other words, is there any additional information contained in our assumptions which would enable us to construct smaller bounds? The answer is "no." This is summarized in Proposition 2. A proof can be found in the appendix.

Proposition 2 *The bounds in Proposition 1 are sharp in the sense that we can choose any point $Z \in [L(W_t), U(W_t)]$ and provide a process for $N(s)$ that satisfies W1 through W3 such that $E[N_t^*|W_t] = Z$.*

4   Bounds with Stronger Assumptions

We now use a stronger set of assumptions to derive tighter bounds. To simplify the discussion, we consider the case where the only source of measurement error is migration. We assume that

$$N(s) = \begin{array}{l} \max\{N_t + j, 1\} \text{ for } s \in [t-1, t-1+\pi_j] \\ N_t \text{ for } s \in (t-1+\pi_j, t] \end{array} \quad \text{for } \Delta M_t = j \text{ and } 0 \leq \pi_j \leq p_j \leq 1 \quad \text{(S)}$$

where p_j is known, but π_j is unknown. This condition places $N(s)$ at either boundary of the intervals in W1 through W3 for fractions of time equal to π_j and $1 - \pi_j$. This gives us that

$$N_t^* = \pi_j * \max\{N_t + j, 1\} + (1 - \pi_j) * N_t \text{ for } \Delta M_t = j. \tag{11}$$

Clearly, if we knew π_j for all j, we could point-identify $E[N_t^*|W_t]$, but we concede that this is very unreasonable. A slightly more realistic situation might be that π_j is unknown, but p_j is known. This might be the case if, for example, a drought which induced a large out-migration occurred during the first half of the survey year. In such a scenario, the researcher might assume that $p_j = \frac{1}{2}$ for all j. Following steps which are similar to those above, we can derive a new set of bounds on $E[N_t^*|W_t]$. The result is summarized in Proposition 3.

Proposition 3 *Assume that the only source of demographic change is migration. Then, if the process for $N(s)$ satisfies S and the probabilities p_j are known for all j, we will have that*

$$\widetilde{L}(W_t) \leq E[N_t^*|W_t] \leq \widetilde{U}(W_t)$$

where

$$\widetilde{U}(W_t) \equiv N_t + \sum_{j>0} j * p_j * P(\Delta M_t = j|W_t)$$

and

$$\widetilde{L}(W_t) \equiv N_t + \sum_{j<0} \max\{p_j * j, 1 - N_t\} * P(\Delta M_t = j|W_t).$$

A few points are worth noting. First, if $p_j < 1$ for some j, then these bounds will be smaller than those in Proposition 1. Second, if $p_j = 1$ for all j, then the bounds in Propositions 1 and 2 will be the same. However, this is not surprising since, if we leave π_j unconstrained in the unit interval, we can construct the same values of N_t^* as we would if we only assumed W1 through W3.

While these bounds are narrower than those in the previous section, they are not that useful because they require the very strong assumption that all migration for all households took place

during a specific time interval. This is almost certain to fail in the vast majority of applications. Consequently, the empirical section of this paper employs the bounds that were derived using the weaker set of assumptions.

5   The Data

Our main data source is the BASIS panel which was administered by the Fundación Salvadoreña para el Desarollo Económico y Social (FUSADES) and The Ohio State University. These data are a sample of rural dwellers. We primarily use the 1999 and 2001 waves of the panel and we employ data on household size as well as the number of migrants and infants (*i.e.* children under 12 months) in the household.[3] Because these data are longitudinal, we can measure migration by taking the difference in the household's migrant stock across survey years. This enables us to calculate the bounds in Section 3. Descriptive statistics for these data can be found in Table 1.

According to researchers at The Ohio State University, the BASIS survey has a stratified design with two strata: households with land and households without land. The sample sizes within strata were determined according to the 1992 census so as to (hopefully) ensure a representative sample.[4] Consequently, no weighting scheme should be necessary. To the best of our knowledge, the survey contains no cluster design. However, we acknowledge the possibility that the observations in the sample are not independent of one another, particularly within small

---

[3]We also used the 1997 wave to allow us to measure migration between 1997 and 1999.

[4]Whether or not the population numbers from the 1992 census are still correct is an open question. Nevertheless, if these numbers are incorrect, aside from running a new census, there is little that we can do to determine the correct numbers.

geographic units. Accordingly, we use the bootstrap to address any possible issues with the survey design. Additional detail about this procedure can be found the next section.

In addition, we use the *Encuesta de Hogares Propósitos Múltiples* (EHPM) which is a consumption survey that is administered annually by the Salvadoran Economic Ministry. In contrast to the BASIS data, this survey covers both rural and urban households. We use a total of 11696 households from the 2001 survey. These data are used to discuss the impact of mismeasured household size on the identification of economies of scale within the household. Summary statistics from the EHPM can also be found in Table 1. Additional detail on the consumption expenditure data can be found in Appendix 2.

The EHPM has a complex two-stage survey design. In the first stage, the country is divided into geographical strata. The Salvadoran Economic Ministry used the 1992 census to determine sample sizes within strata. In the second stage, primary sampling units or clusters were sampled within each strata. Because it is likely that observations within clusters will be correlated, it will also be necessary to adjust all standard errors when working with the EHPM.

6 Estimation and Inference

We use two methods to estimate the bounds in Section 3. The first method is the most straightforward and involves using the BASIS data to estimate the probabilities, $P(\Delta M_t = j | W_t)$, with ordered logit models. We include dummy variables for the household size as well as the number of migrants and infants in the household.[5] These fitted probabilities are then used to back out $U(W_t)$ and $L(W_t)$. One of the advantages of the ordered logit model is that it is easy to

---

[5] We do not address mortality as the BASIS data do not have adequate information on it.

implement. Furthermore, the use of ancillary parameters for each migration category provides us with a flexible way of treating the regression function.[6] One of the disadvantages, however, is that it assumes the size and number of the ancillary parameters are the same for households of all sizes. This is potentially undesirable because it can produce positive probabilities of large positive values of ΔM_t for large households and large negative values for small households. In practice, however, these probabilities are typically small.

Nevertheless, to address this issue, we employ a simple alternative method where we split the sample into households with five or fewer members and households with more than five members and estimate the ordered logits separately for each sample. Doing this mitigates the problem of predicting large positive (negative) values of ΔM_t for larger (smaller) households since the procedure allows the ancillary parameters to vary in size and number with the household's size. After estimating the ordered logits on the split sample, we back out the migration probabilities and calculate the bounds just as before. While this method allows for a more flexible parameterization of the regression function, it should be less efficient than the previous method.[7]

We calculate the standard errors using the bootstrap.[8] We do so for two reasons. First, calculating the analytical standard errors for these bounds is a rather cumbersome task due to

---

[6] We do not use non-parametric estimation due to small sample sizes within the "bins."

[7] We did not further sub-divide the sample into smaller sub-samples, however. The reason for this is that, doing so, involved estimating the ordered logits on rather small sub-samples of the data. These small samples sometimes resulted in non-convergence of the non-linear maximization routine when we bootstrapped our standard errors and, therefore, created substantial complications.

[8] The bootstrapping procedure that we employ works as follows. First, we re-sampled from the data with replacement. To address the possibility of spatial correlation across households, we re-sampled *municipios* from the BASIS data. We re-sampled as many *municipios* as were present in the data. It is unclear from the survey's documentation and our communication with the Ohio State University whether or not the survey had a cluster design. Nevertheless, to the extent that there is spatial correlation across households in these data, our calculation of the standard errors will address it provided that there is only correlation across observations within *municipios*. Using the re-sampled data, we then calculated the bounds. After this, we re-sampled from the data again and repeated the process. After 500 replications, we calculated the standard errors of our estimated bounds.

the large number of ancillary parameters that are being estimated. Second, bootstrapping allows us to address any issues concerning the (possibly) complex design of the survey. As pointed out by Deaton (1997), the bootstrap provides researchers with a convenient, albeit computationally intensive, means of addressing complex survey designs.

To allow us to make inferences about the unidentified parameter, $E[N_t^*|W_t]$, we construct confidence bands which were developed by Imbens and Manski (2004). The confidence intervals that we report cover $E[N_t^*|W_t]$ with at least 95% probability. Note that this is fundamentally different than covering the identified set, $(L(W_t), U(W_t))$, with 95% probability. In general, the intervals that cover the identified set will be larger than those that cover the unidentified parameter and, thus, the confidence intervals that we report should be viewed as conservative in the sense that they will tend to understate any problems associated with measurement errors. The confidence intervals that we report are

$$CI_{0.95} = \left[\widehat{L(W_t)} - C\widehat{\sigma}_{SE,L}, \widehat{U(W_t)} + C\widehat{\sigma}_{SE,U} \right] \tag{12}$$

where $\widehat{\sigma}_{SE,L}$ and $\widehat{\sigma}_{SE,U}$ are the respective standard errors of $\widehat{L(W_t)}$ and $\widehat{U(W_t)}$ and C satisfies

$$\Phi \left(C + \frac{\widehat{U(W_t)} - \widehat{L(W_t)}}{\max \langle \widehat{\sigma}_{SE,L}, \widehat{\sigma}_{SE,U} \rangle} \right) - \Phi \left(-C \right) = 0.95 \tag{13}$$

where $\Phi(.)$ is the CDF of a standard normal random variable.[9] Note that equation (13) has no analytical solution. However, it is trivial to employ numerical methods to calculate C. These

---

[9]Some readers may note that our confidence intervals appear to be slightly different from those in the Imbens and Manski paper. This is because their intervals are defined in terms of the standard deviations of the estimated bounds, whereas ours are defined in terms of the standard errors.

confidence intervals have the desirable property that the probability that they cover $E[N_t^*|W_t]$ will converge uniformly to 95%.

7   Empirical Results

In this section, we discuss our results. Tables 2 and 3 report the estimated bounds and the 95% confidence intervals for the unidentified parameters. Table 2 reports the results for households that have no migrants and one migrant and Table 3 reports the results for households with two migrants and three or more migrants. In both tables, we only report the results for households with no infants. We report the results using both methods for estimating the bounds described in the previous section. We call the first, *Method 1*, and the second, *Method 2*. Finally, we graph the bound estimates in Figures 3 through 10. Figures 3 through 6 use *Method 1* and Figures 7 through 10 use *Method 2*.

The evidence suggests, not surprisingly, that the width of these bounds has a lot to do with the number of migrants in the household. In Figures 3 and 7, we see that the bounds are quite narrow for households that contain no migrants. The results in Table 3 show that the width of the confidence intervals for these households is on the order of 0.30 household members. Moving to households with one migrant each in Figures 4 and 8, we see that the bounds are wider. Calculations in Table 3 show that the width of these confidence intervals is somewhere between 0.60 and 0.90 household members. When we look at households with two migrants in Figures 5 and 9 and three or more migrants in Figures 6 and 10, we see that the situation gets much worse. In Table 3, the width of the confidence intervals is on the order of 1.5 people for households with 2 migrants and 2.8 people for households with 3 or more migrants. Our results were not affected

by our choice of *Method 1* or *Method 2*.

The implications of these findings are significant in our data. Over 30% of the households in our data report having at least one migrant and over 15% report having at least two migrants. Consequently, these bounds are quite wide for a large proportion of the households in our data. We must qualify, once again, that these bounds are necessary but not sufficient for the existence of a large degree of measurement error in household size. Nevertheless, we believe the facts that these bounds are sharp and that they can be so wide suggest that measurement errors in household demographic variables are likely to be a large problem.

8   Implications for Identifying Economies of Scale

We conclude this paper by exploring the implications of mismeasured household size for the identification of economies of scale within the household. To do this, we consider an Engel curve of the form:

$$\omega_f = \alpha + \beta(x - n^*) + \gamma n^* + \varepsilon \tag{14}$$

where ω_f is the share of food in the household's budget, x is the log of total consumption expenditures over the survey period and n^* is the log of N^* where N^* is average household size. We assume that the residual in this equation is uncorrelated with all of the right-hand side regressors. For the sake of simplicity, throughout this section, we suppress all subscripts. This specification was first estimated by Working (1943) and has been used extensively in the literature on household consumer behavior.[10]   As pointed by Deaton (1997) and Deaton and

---

[10]See Lanjouw and Ravallion (1995), Deaton and Paxson (1998) and Deaton and Muellbauer (1986) for some examples.

Muellbauer (1980), this Engel curve has the advantage that it fits the data well and is consistent with optimizing household behavior.

Arguments put forth in a seminal piece by Deaton and Paxson (1998) suggest that γ is a measure of economies of scale within the household and should be positive in most circumstances. The foundation of their argument is that public goods within the household become cheaper as the household's size increases and, if we hold the household's *per capita* expenditures constant, this effectively makes the household richer. To better understand this consider a situation, discussed in Deaton and Paxson's original paper, in which two people decide to move in together. Once these people are living under one roof, they no longer need to pay two separate rents. Provided that their incomes remain constant, each individual has in effect become richer.

Deaton and Paxson go on to argue that if the income elasticity of food is sufficiently high, as it is in the developing world, the household's consumption of food should increase and we should expect to see that γ is positive. However, using data from a variety of countries which run the whole gamut of living standards, they show that the share of food in the household's budget actually *decreases* with household size holding *per capita* expenditures constant. This is the exact opposite of what the theory predicts. The authors consider numerous explanations for their puzzling finding but are ultimately unable to resolve the paradox.[11]

To better understand the role that mismeasured household size can play in the identification

---

[11] In a comment on Deaton and Paxson (1998), Gan and Vernon (2003) claim to resolve the puzzle. The crux of their argument is that there may be relatively large economies of scale in food consumption and, consequently, it may be reasonable to see that the share of food expenditures in the household's budget decreases with household size. The main reason underlying this assertion is that total household expenditures may include goods that are potentially more private than food such as clothes. Gan and Vernon provide evidence that as the household's size rises, food expenditures as a share of food and housing expenditures also rise. They claim that this resolves the puzzle since housing is known to be more public than food. However, Deaton and Paxson (2003), in a response to the comment, assert that Gan and Vernon's findings are consistent with empirical results in their original piece, but do nothing to resolve the puzzle. Their fundamental contention with Gan and Vernon's comment is that it provides little evidence that there are substantial economies of scale in food consumption.

of economies of scale, we first note that, because the household's size is measured with error, equation (14) cannot be estimated since n^* is never observed. Instead, researchers have to estimate

$$\omega_f = \alpha + \beta(x - n) + \gamma n + \upsilon \tag{15}$$

where $n = n^* - e$ and $\upsilon = \varepsilon + (\gamma - \beta)e$. Clearly, OLS will not yield consistent estimates of β and γ since υ is correlated with n. Next, we project e onto x and n and obtain

$$e = \kappa + \phi x + \lambda n + u \tag{16}$$

where u is uncorrelated with both x and n. Because $n = n^* - e$, it is reasonable to expect that $\lambda < 0$ since the covariance between n^* and e will be given by $\sigma_{ne} = \sigma_{n^*e} - \sigma_e^2$. In fact, if the measurement errors are classical in the sense that they are uncorrelated with the true value of the household's size, we will have $\sigma_{ne} = -\sigma_e^2 < 0$. Next, we substitute equation (16) into equation (15) and we obtain

$$\omega_f = \widetilde{\alpha} + \widetilde{\beta}x + \widetilde{\gamma}n + \widetilde{\upsilon} \tag{17}$$

where $\widetilde{\alpha} \equiv \alpha + (\gamma - \beta)\kappa$, $\widetilde{\beta} \equiv \beta + (\gamma - \beta)\phi$, $\widetilde{\gamma} \equiv \gamma + (\gamma - \beta)\lambda$ and $\widetilde{\upsilon} \equiv \varepsilon + (\gamma - \beta)u$.

The probability limit of the OLS estimate of the economies of scale parameter is $\widetilde{\gamma}$. Accordingly, we can write

$$p \lim \widehat{\widetilde{\gamma}} = (1 + \lambda)\gamma - \lambda\beta. \tag{18}$$

This equation illustrates how mismeasured household size can lead to a failure to identify economies of scale even when they are present. To better see this, first note that if λ is

negative, the first term on the right-hand side of the equation will be less than γ and, perhaps, even negative. Second, Engel's Law says that the share of food in the household's budget will fall as the household becomes richer and, thus, β will be negative. Indeed, in nearly every study of household consumption behavior involving Working's Engel curve, estimates of β are always negative and very large. Accordingly, if λ is negative, the second term in the probability limit will be negative and potentially large. In this case, tests for the presence of economies of scale of this type may have low power due to the presence of measurement error in the household's size. Moreover, this calculation suggests that negative estimates of γ may occur even when economies of scale are present. Finally, Deaton and Paxson find that their puzzle is deepest (*i.e.* the estimates of γ are the most negative) for the poorest countries. It is interesting that the poorest countries are also likely to be the ones where household demographic structures are the most pliable as is suggested by Figures 1 and 2.

We conclude this paper with some *prima facie* evidence which suggests that OLS estimates of γ are positively related to β as is suggested by equation (18). To do this, we estimate

$$\omega_f^j = \alpha^j + \beta^j x + \gamma^j n + \sum_{k=1}^{K-1} \eta_k^j \frac{N_k}{N} + \upsilon^j \text{ for } j = 1, ..., J. \tag{19}$$

The dependent variable in this equation is the budget share of a particular food item. The food items that we use are tortillas, bread, rice, milk, beans, chicken, beef, pork, vegetables, fruit and eggs. The term $\frac{N_k}{N}$ is the share of the total number of household members in a particular age and gender category. We report the estimates of γ^j and β^j in Table 4. What can be seen in the table is that the estimates of γ^j are related to the estimates of β^j. Generally, we see that food items with higher income elasticities also have higher estimates of γ^j. To better see this,

we plot the pairs $\left(\widehat{\gamma^j}, \widehat{\beta^j}\right)$ in Figure 11 which clearly illustrates a strong positive relationship between the two parameter estimates.[12]

The results and calculations of this section suggest that mismeasured household size *may* help to explain the paradox that Deaton and Paxson originally posed. However, we caution that we still do not understand the magnitude (or even the sign) of the parameter λ fully. While we believe that it is reasonable to suspect that λ is negative and potentially large, further work is warranted.

9   Conclusions

In this paper, we investigated the idea that demographic change can induce errors in the measurement of the household's composition. To do this, we began by formally defining what it means to mismeasure the household's size. We then derived bounds on the household's size using weak assumptions on demographic dynamics during the survey period. Estimation using Salvadoran survey data revealed that these bounds are often wide and, thus, suggested that there may be a lot of measurement error in variables on household composition. However, we

---

[12]There are two alternative explanations for the positive relationship in Figure 11. The first is that goods that have higher income elasticities also have fewer economies of scale associated with them than the other goods in the household's budget. If this were, in fact, the case, then we would see that, as the household's size increases, the prices of the other goods in the budget would decrease more rapidly than the goods with the higher income elasticities. However, if this were true, then these results suggest that there are fewer economies of scale in beef consumption than in pork consumption. It is unclear to us why this would be the case. The second explanation for the relationship in the figure has to do with the theory in Deaton and Paxson's original work. Specifically, they show that the consumption of a good should increase with the household's size when the income elasticity of that good is high relative to the absolute value of its price elasticity. The fact that we find positive estimates of the economies of scale parameter for goods that are luxuries (or almost luxuries) like beef or pork suggests that there may be some credence to this. However, working against this explanation is the presumption that the price elasticity of beef or pork is higher than the price elasticity of staples like tortillas. Unfortunately, without data on unit prices, there is no way of verifying this presumption. In addition, this argument suggests that the negative estimates of the economies of scale parameter for tortillas is the result of the absolute value of the price elasticity of a staple being high relative to its income elasticity which we find to be somewhat hard to believe.

are careful to point out that the evidence that we present is necessary but not sufficient for the existence of a large degree of measurement error in demographic variables. We concluded by showing how mismeasured household size could potentially lead to a failure to identify economies of scale within the household even when they are present.

This work suggests several avenues for future research. First, an informative exercise would be to calculate similar bounds using survey data from other countries. We conjecture that these bounds will be narrower for richer countries where household structures are more likely to be more stable over time. Second, more work is needed to better understand implications for the identification of economies of scale and, more generally, for the estimation of Engel curves. Simple calculations suggest that measurement errors in household size may possibly shed some light on the paradox posed in Deaton and Paxson (1998). However, we caution the reader that we view these results as preliminary. Finally, it would be interesting to derive bounds on the parameters of the Engel curve in equation (14) when household size is mismeasured.[13]

10   Appendix 1 - Proof of Proposition 2

Proof. The goal of this proof is to produce a set of processes for $N(s)$ for all households that satisfy assumptions W1 through W3 such that $Z = E[N_t^*|W_t]$ for $Z \in [L_t(W_t), U(W_t)]$. For the sake of simplicity, we consider the case where the only source of demographic change is migration.

---

[13]Hu (2005) derives bounds on the parameters of a linear regression model when one of the regressors contains non-classical measurement error. However, his methodology is not applicable to our case since it requires observation of the true regressor (*i.e.* the regressor without any measurement error) from another data set.

We begin by writing

$$E[N_t^*|W_t] = e_{-J}*p_{-J}+...+e_{-1}*p_{-1}+e_0*(1-p_{-J}-...-p_{-1}-p_1-...p_J)+e_1*p_1+...+e_J*p_J \quad (20)$$

where $e_i \equiv E[N_t^*|\Delta M_t = i, W_t]$ and $p_i \equiv P(\Delta M_t = i|W_t)$. Next, we consider the case where $Z \in [U_t(W_t) - p_J J, U(W_t)]$. We now choose the following processes for $N(s)$:

$$N(s) = \begin{cases} N_t \text{ for } s \in (t-1,t] \\ N_t + j \text{ for } s = t-1 \end{cases} \quad \text{and } \Delta M_t = j \leq 0 \quad (21)$$

$$N(s) = \begin{cases} N_t \text{ for } s = t \\ N_t + j \text{ for } s \in [t-1,t) \end{cases} \quad \text{and } 0 < \Delta M_t = j < J \quad (22)$$

We assume that these two conditions hold for *all* households. These conditions assume that, for $\Delta M_t < J$, the household's size is constant over a time interval of measure one. Clearly, these conditions satisfy W1 through W3. These conditions on the $N(s)$ process then give us that

$$E[N_t^*|W_t] = N_t * (1 - p_J) + p_1 + 2p_2 + ... + (J-1)p_{J-1} + p_J * e_J \quad (23)$$

since they hold for all households. If we set the above expression equal to Z, we can then write e_J as

$$e_J = N_t + p_J^{-1} * [Z - p_1 - ... - (J-1)p_{J-1} - N_t]. \quad (24)$$

Next, noting that

$$U(W_t) = N_t + p_1 + ... + Jp_J, \quad (25)$$

23

and recalling that $Z \in [U_t(W_t) - p_J J, U(W_t)]$, we will have that

$$e_J = E[N_t^* | \Delta M_t = J, W_t] \in [N_t, N_t + J]. \tag{26}$$

Finally, we choose

$$N(s) = \begin{array}{l} N_t \text{ for } s \in (t - \delta, t] \\ \\ N_t + J \text{ for } s \in [t - 1, t - \delta] \end{array} \qquad \text{for } \Delta M_t = J \tag{27}$$

where $\delta \equiv 1 + \frac{N_t - e_J}{J}$ where e_J is defined as in equation (24). Note that $\delta \in [0, 1]$ since $e_J \in [N_t, N_t + J]$. The proof for the other values of Z is completely analogous. ∎

11 Appendix 2 - Consumption Expenditures in the EHPM

The EHPM contains detailed information on consumption expenditures which is summarized in Table 5. The data on food expenditures as well all expenditures in consumption categories 1 and 2 includes all items purchased on the market, produced at home and received as aid. Total consumption is the sum of all expenditures in categories 1 and 2 plus expenditures on food, utilities, schooling and medical care. We did not include expenditures on housing as these data were suspect.[14]

---

[14]Discussions with a researcher at FUSADES, a Salvadoran think tank, corroborated these suspicions. In addition, it is important to note that the lack of data on housing expenditures does not impact the analysis in Section 8. The reason is that our analysis focuses on the biases that measurement error can create when estimating Engel curves. These biases will be present with or without the housing expenditure data.

References

[1] Deaton, Angus. 1997. *The Analysis of Household Surveys.* Baltimore: Johns Hopkins University Press.

[2] Deaton, Angus and John Muellbauer. 1980. *Economics and Consumer Behavior.* Cambridge: Cambridge University Press.

[3] Deaton, Angus and John Muellbauer. 1986. "On Measuring Child Costs: with Application to Poor Countries." *Journal of Political Economy.* 94: 720-744.

[4] Deaton, Angus and Christina Paxson. 1998. "Economies of Scale, Household Size and the Demand for Food." *Journal of Political Economy.* 106: 897-930.

[5] Deaton, Angus and Christina Paxson. 2003. "Engel's What? A Response to Gan and Vernon." *Journal of Political Economy.* 111: 1378-1381.

[6] Gan, Li and Victoria Vernon. 2003. "Testing the Barten Model of Economies of Scale in Household Consumption: Toward Resolving a Paradox of Deaton and Paxson." *Journal of Political Economy.* 111: 1361-1377.

[7] Hu, Yingyao. 2005. "Bounding Parameters in a Linear Regression Model with a Mismeasured Regressor using Additional Information." *Journal of Econometrics.* 133: 51-70.

[8] Imbens, Guido and Manski, Charles. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica,* 72: 1845 - 1857.

[9] Lanjouw, Peter and Ravallion, Martin. 1995. "Poverty and Household Size," *The Economic Journal,* 105: 1415-1434.

[10] Working, Holbrook. 1943. "Statistical Laws of Family Expenditure," *Journal of the American Statistical Association*. 104: 275-298.
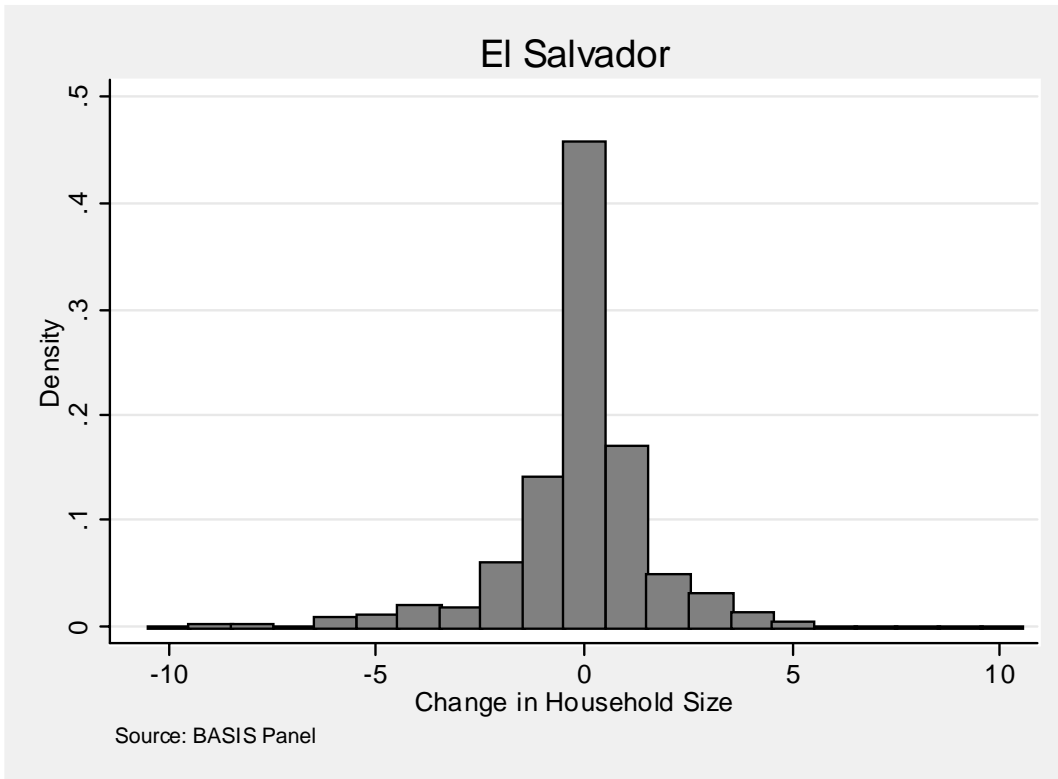
Figure 1

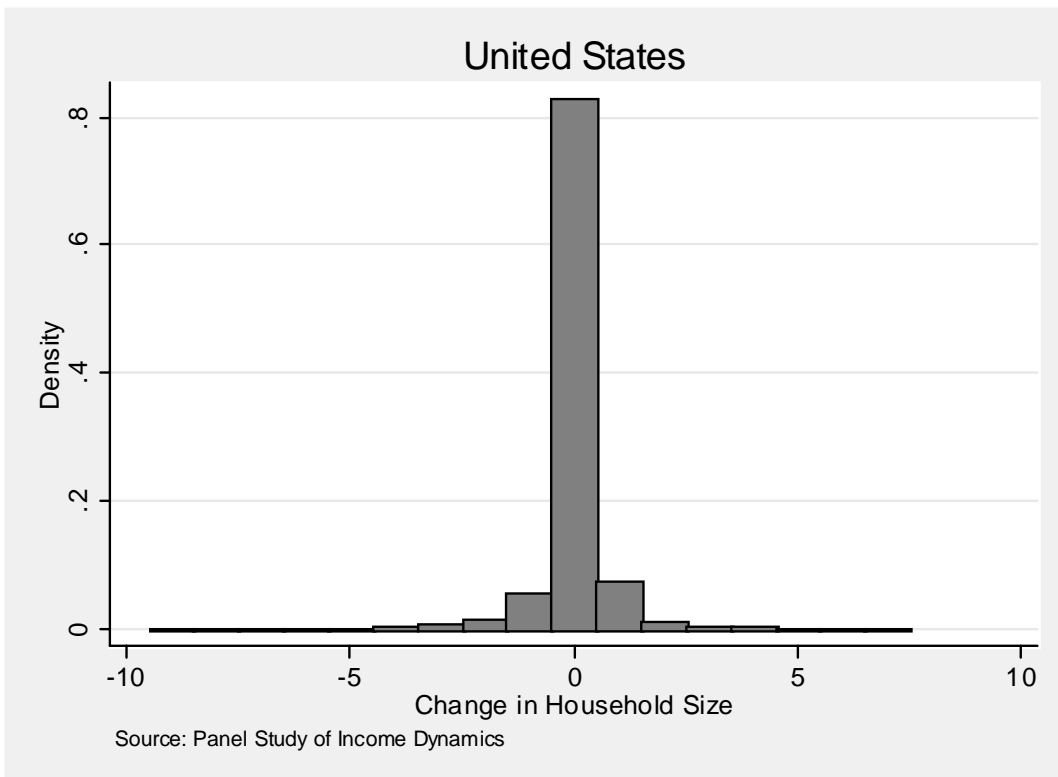El Salvador

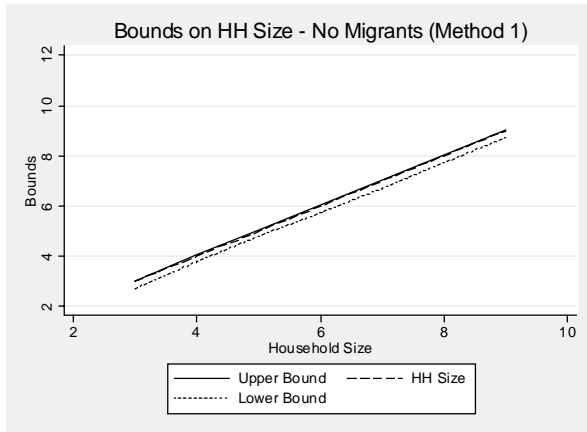Source: BASIS Panel

Figure 2

United States

Source: Panel Study of Income Dynamics

Figure 3

Bounds on HH Size - No Migrants (Method 1)

Figure 5

Bounds on HH Size - Two Migrants (Method 1)

Figure 4

Bounds on HH Size - One Migrant (Method 1)

Figure 6

Bounds on HH Size - Three Migrants (Method 1)

Figure 7

Bounds on HH Size - No Migrants (Method 2)

Figure 9

Bounds on HH Size - Two Migrants (Method 2)

Figure 8

Bounds on HH Size - One Migrant (Method 2)

Figure 10

Bounds on HH Size - Three Migrants (Method 2)

Figure 11

Economies of Scale and the Income Elasticity of Food
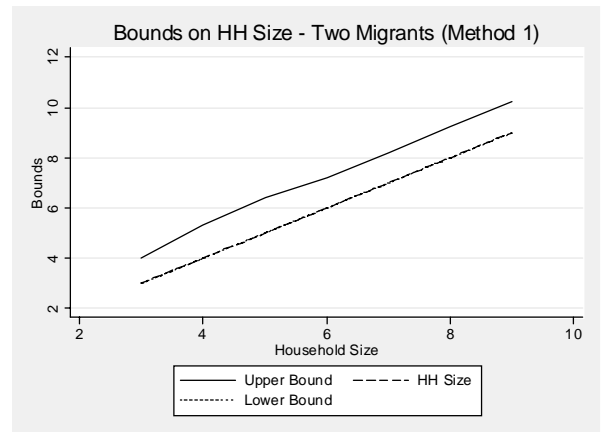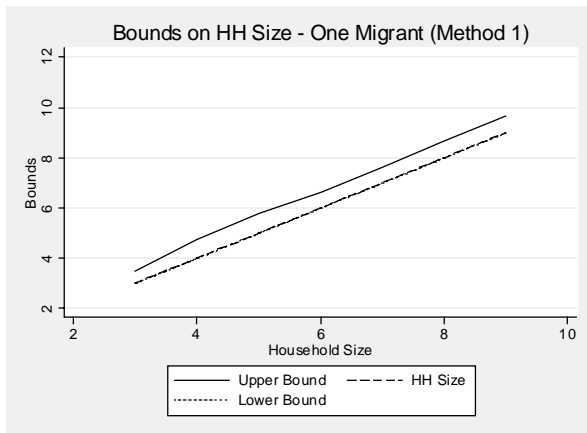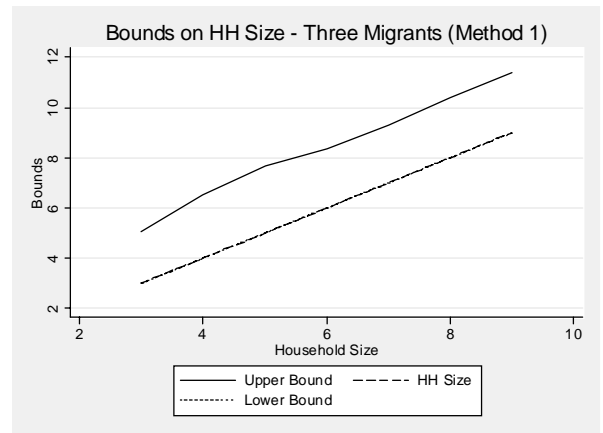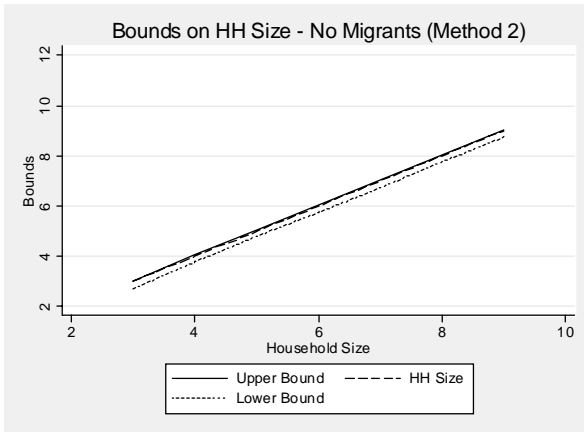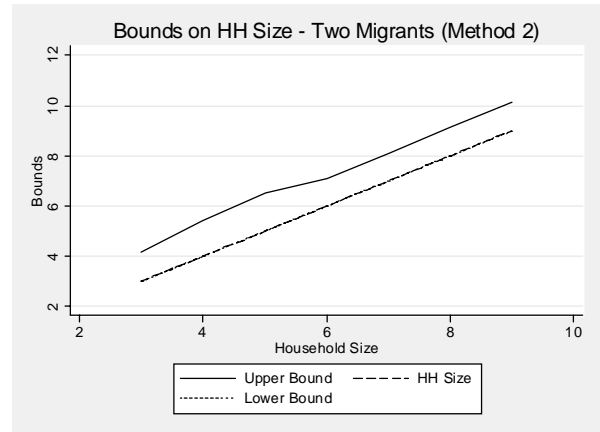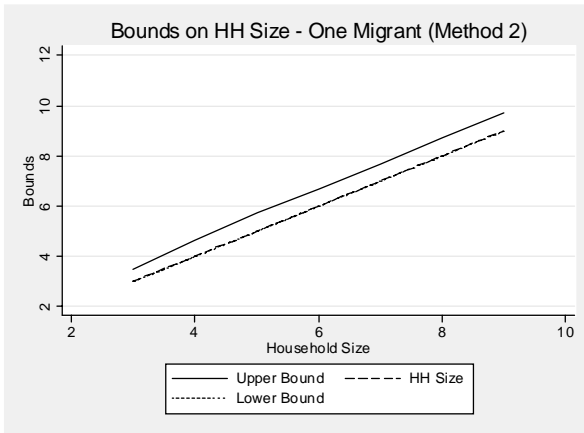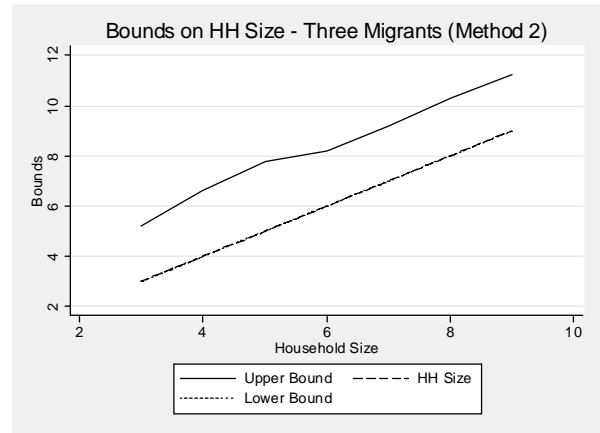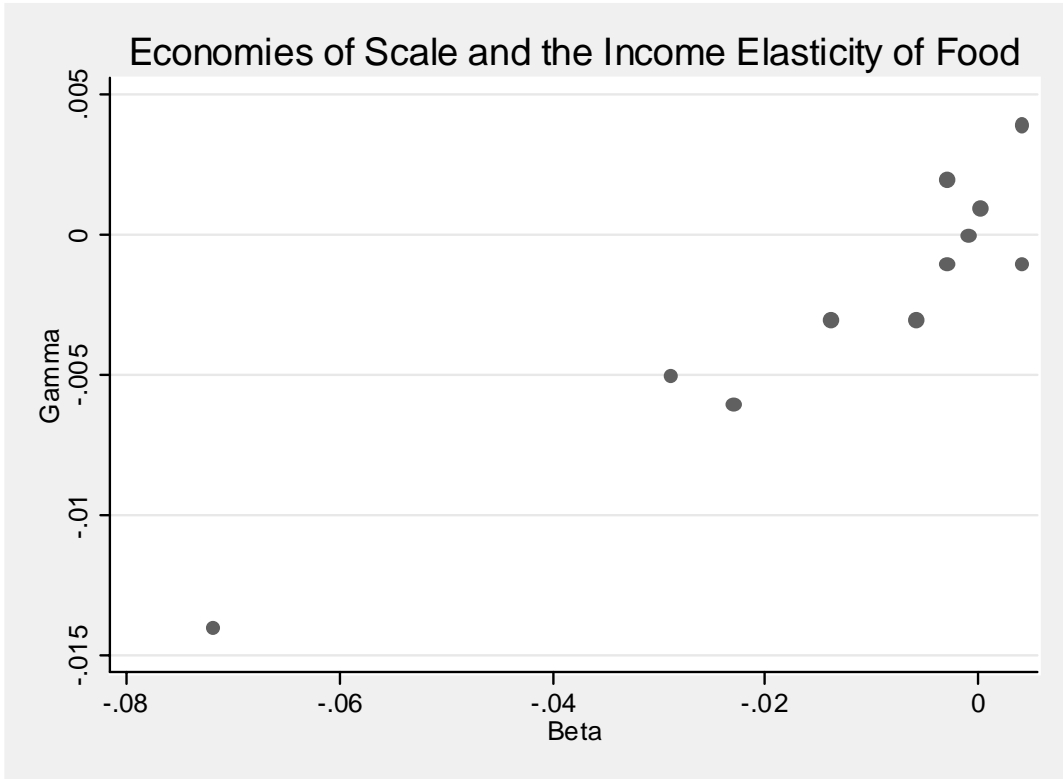
Table 1: Descriptive Statistics

	Mean (Standard Deviation)
BASIS[1]	
Household Size	5.96 (2.68)
Migrants	0.64 (1.32)
Infants	0.06 (0.25)
EHPM[2]	
Total Consumption Expenditures	3044.98 (2223.48)
Household Size	4.44 (2.26)

[1]The sample size for these data is 1265 households.
[2]The sample size for these data is 11696.

Table 2: Bounds - Zero or One Migrant, No Infants

HH Size*	$\left(\widehat{L(W_t)}, \widehat{U(W_t)}\right)$	95% CI for $E[N_t^*\|W_t]$	$\left(\widehat{L(W_t)}, \widehat{U(W_t)}\right)$	95% CI for $E[N_t^*\|W_t]$
	Method 1		Method 2	
No Migrants				
3	(2.708, 3.029)	[2.633, 3.043]	(2.702, 3.030)	[2.615, 3.047]
4	(3.788, 4.049)	[3.728, 4.071]	(3.777, 4.050)	[3.711, 4.077]
5	(4.808, 5.059)	[4.752, 5.082]	(4.794, 5.060)	[4.723, 5.087]
6	(5.725, 6.042)	[5.646, 6.062]	(5.747, 6.042)	[5.672, 6.060]
7	(6.710, 7.041)	[6.628, 7.057]	(6.731, 7.040)	[6.657, 7.057]
8	(7.733, 8.045)	[7.636, 8.065]	(7.756, 8.044)	[7.682, 8.062]
9	(8.726, 9.044)	[8.632, 9.070]	(8.745, 9.043)	[8.668, 9.060]
One Migrant				
3	(2.986, 3.502)	[2.975, 3.613]	(2.983, 3.463)	[2.967, 3.576]
4	(3.990, 4.711)	[3.982, 4.883]	(3.988, 4.650)	[3.976, 4.822]
5	(4.992, 5.792)	[4.985, 5.944]	(4.990, 5.724)	[4.978, 5.903]
6	(5.988, 6.640)	[5.977, 6.791]	(5.990, 6.705)	[5.980, 6.864]
7	(6.987, 7.625)	[6.976, 7.751]	(6.989, 7.684)	[6.979, 7.833]
8	(7.988, 8.667)	[7.977.8.833]	(7.990, 8.736)	[7.980, 8.878]
9	(8.988, 9.662)	[8.978, 9.822]	(8.990, 9.717)	[8.979, 9.856]

*Refers to the household's reported size at the time of the survey.

Table 3: Bounds - Two or More than Three Migrants, No Infants

HH Size*	$\left(\widehat{L(W_t)}, \widehat{U(W_t)}\right)$	95% CI for $E[N_t^*\|W_t]$	$\left(\widehat{L(W_t)}, \widehat{U(W_t)}\right)$	95% CI for $E[N_t^*\|W_t]$
	Method 1		Method 2	
	Two Migrants			
3	$(2.996, 4.008)$	$[2.991, 4.229]$	$(2.997, 4.144)$	$[2.992, 4.474]$
4	$(3.997, 5.304)$	$[3.994, 5.557]$	$(3.998, 5.431)$	$[3.995, 5.772]$
5	$(4.997, 6.415)$	$[4.995, 6.638]$	$(4.998, 6.539)$	$[4.995, 6.848]$
6	$(5.996, 7.206)$	$[5.993, 7.436]$	$(5.995, 7.104)$	$[5.990, 7.363]$
7	$(6.996, 8.185)$	$[6.992, 8.400]$	$(6.995, 8.077)$	$[6.990, 8.325]$
8	$(7.996, 9.244)$	$[7.992, 9.511]$	$(7.996, 9.145)$	$[7.990, 9.387]$
9	$(8.996, 10.236)$	$[8.993, 10.483]$	$(8.995, 10.120)$	$[8.990, 10.360]$
	Three or More Migrants			
3	$(2.999, 5.053)$	$[2.998, 5.470]$	$(2.999, 5.204)$	$[2.999, 5.708]$
4	$(3.999, 6.505)$	$[3.999, 6.969]$	$(3.999, 6.635)$	$[3.999, 7.187]$
5	$(4.999, 7.675)$	$[4.999, 8.120]$	$(4.999, 7.799)$	$[4.999, 8.313]$
6	$(5.999, 8.355)$	$[5.999, 8.831]$	$(5.999, 8.223)$	$[5.998, 8.759]$
7	$(6.999, 9.323)$	$[6.999, 9.750]$	$(6.999, 9.181)$	$[6.998, 9.714]$
8	$(7.999, 10.413)$	$[7.999, 10.945]$	$(7.999, 10.286)$	$[7.998, 10.820]$
9	$(8.999, 11.401)$	$[8.999, 11.965]$	$(8.999, 11.248)$	$[8.998, 11.787]$

*Refers to the household's reported size at the time of the survey.

Table 4: Engel Curve Estimates

	$\widehat{\beta}_j$	$\widehat{\gamma}_j$
Tortillas	−0.072	−0.014
	(−30.71)	(−8.03)
Beans	−0.029	−0.005
	(−26.72)	(−8.20)
Eggs	−0.023	−0.006
	(−31.85)	(−7.85)
Rice	−0.014	−0.003
	(−28.28)	(−7.68)
Vegetables	−0.006	−0.003
	(−9.38)	(−5.34)
Bread	−0.003	−0.001
	(−2.34)	(−0.87)
Chicken	−0.003	0.002
	(−2.91)	(2.26)
Milk	−0.001	−0.000
	(−1.28)	(−0.36)
Pork	0.000	0.001
	(0.91)	(3.37)
Beef	0.004	0.004
	(4.60)	(5.02)
Fruit	0.004	−0.001
	(7.64)	(−1.58)

*This table contains OLS estimates of the Engel curves described in Section 8. All standard errors allow for clustering on municipios.

Table 5: Constituents of Consumption Expenditures: EHPM

Component	Mean (Standard Deviation)	Contents
Food	1423.33 (880.54)	tortillas, bread, rice, beans, salt, sugar, grains, chicken, beef, pork, fish, eggs, milk, cheese, aceite, vegetables, fruits, restaurant meals, prepared meals, coffee, drinks, alcohol, other items
Category 1	244.51 (252.65)	toiletries, soap, cleaning products, magazines, newspapers, cosmetics, fuel, transportation, babysitting
Category 2	167.26 (343.92)	travel, jewelry, pots, towels, car repairs, other repairs, appliances, furniture, clothes, glasses
Utilities	461.41 (541.76)	water, electricity, kerosene, propane, candles, carbon, leña, telephone, cell phone, cable, garbage
School Expenditures	677.80 (984.91)	tuition, supplies, uniforms, textbooks
Medical Expenses	70.68 (320.00)	doctor's visits, lab work, x-rays, hospital days, medicine
Total Expenses	3044.98 (2223.48)	