Appendix:

February 6, 2023

1 Formal Assumptions and Proofs

Here we formally derive our model, and prove that it is semiparametrically point identified. To simplify the derivations and assumptions, we first prove results without unobserved random utility parameters (as would apply if, e.g., our data consisted of many observations of a single household, or of many households with no unobserved variation in tastes). We then later add unobserved error terms to the model, corresponding to unobserved preference heterogeneity.

Let f, r, y, p, π, and z be as defined in the main text. Note that the first few Lemmas below will not impose the restriction that f only equal two values.

ASSUMPTION A1: Conditional on f, r, y, p, π, and z, the household chooses quantities to consume using the program given by equation (6) in the main text.

Assumption A1 describes the collective household's conditionally efficient behavior. For each household member j, U_j is that member's utility function over consumption goods, u_j is that members additional utility or disutility associated with f, and ω_j is that member's Pareto weight.

As can be seen by equation (6) in the main text, the way that private assignable goods q_j differ from other goods g is that each q_j only appears in the utility function of individual j (which makes it assignable to that member) and these goods are unaffected by the matrix A_f in the budget constraints, meaning that they are not shared or consumed jointly (which makes them private goods).

We next assume some regularity conditions. These assumptions ensure sensible and convenient restrictions on economic behavior like no money illusion, preferring larger consumption bundles to smaller ones, and the absence of corner solutions in the household's maximization problem.

ASSUMPTION A2: Each $\omega_j(f, z, p, \pi, y)$ function is differentiable and homogeneous of degree zero in (p, π, y). Each $U_j(q_j, g_j, z)$ function is concave, strictly increasing, and twice continously differentiable in g_j and q_j. For each f, the matrix A_f is nonsingular with all nonnegative elements and a strictly positive diagonal. The variable y and each element of p and π are all strictly positive, and the maximizing values of $g_1, q_1, ...g_J, q_J$ in Assumption A1 are all strictly positive.

LEMMA 1. Let Assumptions A1 and A2 hold. Then there exist positive resource share functions $\eta_j(p,\pi,y,f,z)$ such that $\sum_{j=1}^{J}\eta_j(p,\pi,y,f,z)=1$, and the household's demand function for goods is given by each member j solving the program

$$\max_{g_j,q_j} U_j(q_j,g_j,z) \tag{1}$$

$$\text{such that } \; p'A_f g_j+\pi_j q_j=\sum\nolimits_{j=1}^{J}\eta_j(p,\pi,y,z,f)\,y \; \text{ and } g=A_f\sum\nolimits_{j=1}^{J}g_j.$$

To prove Lemma 1, first observe that the values of $g_1,q_1,...g_J,q_J$ that maximize equation (6) in the main text are equivalent to the values that maximize

$$\max_{g_1,q_1,...g_J,q_J}\sum\nolimits_{j=1}^{J}U_j(q_j,g_j,z)\,\omega_j(p,\pi,y,f) \tag{2}$$

given the same budget constraint. because the terms in equation (6) in the main text that are not in (2) do not depend on $g_1,q_1,...g_J,q_J$. With that replacement, the proof of Lemma 1 then follows immediately from the results derived in BCL. BCL only considered $J=2$, but the extension of this Lemma to more than two household members, and to carrying the additional covariates, is straightforward. Note that the resource share functions η_j in Lemma 1 do not depend on r, because r, including the component v, does not appear in either equation (2) or in the budget constraint, and so cannot affect the outcome quantities.

Our empirical work will make use of cross section data, where price variation is not observed. Most of the remaining assumptions we make about resource shares and about the U_j component of utility are the same, or similar, to those made by DLP, and for the same reason: to ensure identification of the model without requiring price variation.

ASSUMPTION A3. The resource share functions $\eta_j(p,\pi,y,f,z)$ do not depend on y.

DLP give many arguments, both theoretical and empirical, supporting the assumption that resource shares do not vary with y. Given Assumption A3, we hereafter write the resource share function as $\eta_j(\pi,p,f,z)$.

For the next assumption, recall that an indirect utility function is defined as the function of prices and the budget that is obtained when one substitutes an individual's demand functions into their direct utility function.

ASSUMPTION A4. For each household member j, the direct utility function $U_j(g_j,q_j,z)$, when facing prices p and π and having the budget y, has the associated indirect utility function

$$V_j(\pi_j,p,y,z)=[\ln y-\ln S_j(\pi_j,p,z)]\,M_j(\pi_j,p,z) \tag{3}$$

For some functions S_j and M_j.

Assumption A4 says that household members each have utility functions in the class that Muellbauer (1974) called PIGLOG (price independent, generalized logarithmic) preferences. As noted in the main text, this is a class of functional forms that is widely known to fit empirical continuous consumer demand data well. Examples of popular models in this class include the Christensen, Jorgenson, and Lau (1975) Translog demand system and Deaton and Muellbauer's (1980) AIDS (Almost Ideal Demand System) model.[1]

LEMMA 2: Let Assumptions A1, A2, A3, and A4 hold. Then the value of $U_j(q_j, g_j, z)$ attained by household member j is given by

$$U_j = [\ln \eta_j(\pi, A_f p, f, z) + \ln y - \ln S_j(\pi_j, A_f p, z)] \, M_j(\pi_j, A_f p, z) \tag{4}$$

and the household's demand functions for the private assignable goods q_j are

$$q_j = \eta_j(\pi, A_f p, f, z) \, y \left(\frac{\partial \ln S_j(\pi_j, A_f p, z)}{\partial \pi_j} - \frac{\partial \ln M_j(\pi_j, A_f p, z)}{\partial \pi_j} \ln \left(\frac{\eta_j(\pi, A_f p, f, z) \, y}{S_j(\pi_j, A_f p, z)} \right) \right) \tag{5}$$

To prove Lemma 2, observe that by Lemma 1, household member j maximizes the utility function $U_j(q_j, g_j, z)$ facing shadow prices $A'_f p$ and π_j and having the shadow budget $\eta_j(\pi, A_f p, f, z) \, y$. Therefore, using the definition of indirect utility, member j's attained utility level $U_j(q_j, g_j, z)$ is given by $V_j\left(\pi_j, A'_f p, \eta_j(\pi, A_f p, f) \, y\right)$, which by Assumption A4 equals equation (4). Next, a property of regular indirect utility functions is that the corresponding demand functions can be obtained by Roy's identity. Equation (5) is obtained by applying Roy's identity to equation (3) for the private assignable goods q_j, and then replacing p and y in the result with $A'_f p$ and $\eta_j(\pi, A_f p, f) \, y$.

We could similarly obtain the demand functions for other goods g, as in BCL, but these will be more complicated due to the sharing, with Roy's identity being applied to each member to obtain each g_j demand function, and substituting the results into $g = A_f \sum_{j=1}^{J} g_j$. However, our empirical analyses will only make use of the private assignable goods q_j with demands given by equation (5).

ASSUMPTION A5. Let $\ln M_j(\pi_j, A_f p, z) = m_j(A_f p, z) - \beta(z) \ln \pi_j$ for some functions m_j and β.

There are two restrictions embodied in Assumption A5. One is that the functional form of $\ln M_j$ in terms of prices is linear and additive in $\ln \pi_j$, and the other is that the function $\beta(z)$ does not vary by j. The functional form restriction of log linearity in log prices is a common one in consumer demand models, e.g., the function M_j in Deaton and Muellbauer's (1980) AIDS (Almost Ideal Demand System) satisfies this restriction. Assumption A5 could be further relaxed by letting β depend on p (though not on A_f) without affecting later results.

To identify their model, DLP define and use a property of preferences called similarity across people (SAP), and provide empirical evidence in support of SAP. The restriction that β not vary by j suffices to make SAP hold for the private assignable goods (but not necessarily for other goods).

[1] Most more recent alternatives, like so-called "rank three" demand systems, are used for data from countries where the distribution of y is large, and more complicated budget responses are needed to capture behavior at both low and high income levels. Other popular demand models, like the multinomial logit based models widely used in the industrial organization literature, are designed for use with discrete demand data and are unsuitable for the type of continuous consumer demand data we analyze here.

ASSUMPTION A6. Let $\ln S_j(\pi_j, A_f p, z) = \ln s_j(\pi_j, p, z) - \ln \delta(A_f p, z)$ for some functions s_j and δ. Without loss of generality, let $\ln \delta(A_0 p, z) = 0$.

Assumption A6 assumes separability of the effects of π_j and f on the function S_j. DLP discuss various ways in which the matrix A_f can drop out of a function of prices, as required in the function s_j.[2] This assumption is not vital, but will be helpful for making the cost of an inefficient choice of f identifiable. Assuming $\ln \delta(A_0 p, z) = 0$ in Assumption A6 is without loss of generality, because if it does not hold then one can make it hold if one redefines δ and s_j by subtracting $\ln \delta(A_0 p, z)$ from both $\ln \delta(f, p, z)$ and $\ln s_j(\pi_j, p, z)$.

It will be convenient to express our demand functions in budget share form. Define $w_j = q_j \pi_j / y$. This budget share is the fraction of the household's budget y that is spent on buying person j's assignable good q_j.

LEMMA 3: Given Assumptions A1 to A6, the value of $U_j(q_j, g_j, z)$ attained by household member j is given by

$$[\ln \eta_j(\pi, A_f p, f, z) + \ln y - \ln s_j(\pi_j, p, z) + \ln \delta(A_f p, z)]\,[m_j(A_f p, z) - \beta(z) \ln \pi_j] \qquad (6)$$

and the budget share demand functions for each private assignable good are given by

$$w_j = \eta_j(\pi, A_f p, f, z)\,[\gamma_j(\pi_j, p, z) + \beta(z)(\ln y + \ln \eta_j(\pi, A_f p, f, z) + \ln \delta(A_f p, z))]\,. \qquad (7)$$

where the function γ_j is defined by

$$\gamma_j(\pi_j, p, z) = \frac{\partial \ln s_j(\pi_j, p, z)}{\partial \ln \pi_j} - \beta(z) \ln s_j(\pi_j, p, z)$$

The proof of Lemma 3 consists of substituting the expressions for M_j and S_j given by Assumptions A5 and A6 into the equations given by Lemma 2, and converting the quantity q_j into the budget share w_j.

ASSUMPTION A7. Market prices p and π are the same for all households.

Our data come from a single time period, which (assuming the law of one price) justifies assuming p and π are the same across all households. This assumption makes our demand functions reduce to Engel curves. For simplicity, we abuse notation here and redefine objects that were functions of $A_f p$ as just functions of f, since with fixed prices the only source of variation of $A_f p$ is just variation in f).

[2]For example, one way A_f drops out is if A_f is block diagonal, with one block that does not vary by f, and with s_j only depending on π_j and the prices in that block. Alternatively, linear constraints could be imposed on the elements of A_f, with s_j depending only on the corresponding functions of prices, that, by these constraints, do not vary with A_f. Analogous restrictions are often imposed on demand systems. For example as shown in Lewbel (1991), the Translog demand system as implemented by Jorgenson, and Slesnick (1987) imposes a linear constraint on its Barten (1964) scales, that results in a restriction like this on its equivalence scales. Note that BCL refer to the diagonal elements of A_f as Barten technology parameters, due to their equivalence to Barten scales.

LEMMA 4: Given Assumptions A1 to A7, the value of $U_j(q_j, g_j, z)$ attained by household member j is given by

$$[\ln \eta_j(f,z) + \ln y - \ln s_j(z) + \ln \delta(f,z)] M_j(f,z) \tag{8}$$

and the budget share Engel curve functions $w_j = W_j(f,z,y)$ for each private assignable good are given by

$$W_j(f,z,y) = \eta_j(f,z)\left[\gamma_j(z) + \beta(z)(\ln y + \ln \eta_j(f,z) + \ln \delta(f,z))\right]. \tag{9}$$

Lemma 4 entails a small abuse of notation, where we have absorbed the values of p and π into the definitions of all of our functions, noting that any function of $A_f p$ remains a function of f even if p is a constant. Lemma 4 is just rewriting Lemma 3 after dropping the prices.

LEMMA 5: Let Assumptions A1 to A7 hold. Let $W_j(f,z,y)$ be defined by equation (9) for $j = 1,...,J$. Given functions $W_j(f,z,y)$, the functions $\eta_j(f,z)$, $\delta(f,z)$,$\gamma_j(z)$, and $\beta(z)$ are all point identified.

To prove Lemma 5, observe first by equation (9) that $\eta_j(f,z)\beta(z) = \partial W_j(f,z,y)/\partial \ln y$. Next, since resource shares sum to one, we can identify $\beta(z)$ and $\eta_j(f,z)$ by

$$\beta(z) = \sum\nolimits_{j=1}^{J} \frac{\partial W_j(f,z,y)}{\partial \ln y} \quad \text{and} \quad \eta_j(f,z) = \frac{1}{\beta(z)} \frac{\partial W_j(f,z,y)}{\partial \ln y}$$

Next, define $\rho_j(f,z,y)$ by

$$\rho_j(f,z,y) = \frac{W_j(f,z,y)}{\eta_j(f,z)} - \beta(z)(\ln y + \ln \eta_j(f,z))$$

The function $\rho_j(f,z,y)$ is identified because it is defined entirely in terms of identified functions. By equation (9), $\rho_j(f,z,y) = \gamma_j(z) - \beta(z)\ln \delta(f,z)$. It follows from Assumption A6 that $\ln \delta(0,z) = 0$, so $\gamma_j(z)$ and $\delta(f,z)$ are identified by

$$\gamma_j(z) = \rho_j(0,z,y) \quad \text{and} \quad \ln \delta(f,z) = \frac{\rho_j(f,z,y) - \rho_j(0,z,y)}{\beta(z)}$$

evaluated at any value of y (or, e.g., averaged over y).

Lemma 5 shows that, given the household demand functions, the resource share functions $\eta_j(f,z)$ are identified, so our model, like DLP, overcomes the problem in the earlier collective household literature of (the levels of) resource shares not being identified. Lemma 5 also shows identification of the preference related functions $\gamma_j(z)$ and $\beta(z)$, and identification of our new cost of inefficiency function $\delta(f,z)$.

LEMMA 6: Let Assumptions A1 to A7 hold. Assume f is determined by maximizing $\Psi(U_1 + u_1, ..., U_J + u_J)$ for some function Ψ. Then $f = \arg\max \Psi(R_1(p,y,f,v), ...R_J(p,y,f,v))$ where $R_j(f,y,v,z)$ is given by

$$R_j(f,y,v,z) = (\ln \eta_j(f,z) + \ln y - \ln s_j(z) + \ln \delta(f,z)) M_j(f,z) + u_j(f,v,z)$$

The proof of Lemma 6 is then that, by equation (8) and the definition of u_j, for any f the level of $U_j + u_j$ attained by member j is given by the function $R_j(f, y, v, z)$.

The above analyses apply to a single household. Our data will actually consist of a cross section of households, each only observed once. To allow for unobserved variation in tastes across households in a conveniently tractible form, replace the function $\ln S_j(\pi_j, A_f p, z)$ with $\ln S_j(\pi_j, A_f p, z) - \widetilde{\varepsilon}_j$ where $\widetilde{\varepsilon}_j$ is a random utility parameter representing unobserved variation in preferences for goods. This means that $\widetilde{\varepsilon}_j$ appears in member j's utility function U_j. We assume these taste parameters vary randomly across households, so $E(\widetilde{\varepsilon}_j \mid r, z) = 0$. Similarly, replace $u_j(f, r, z)$ with $u_j(f, r, z) + \widetilde{e}_{jf}$ where $\widetilde{e}_{jf}$ represents variation in the utility or disutility associated with the choice of f. The errors $\widetilde{e}_{jf}$ and $\widetilde{\varepsilon}_j$ can be correlated with each other and across household members.

Substituting these definitions into the above equations, we get

$$w_j = \eta_j(f, z)\left[\gamma_j(z) + \beta(z)(\ln y + \ln \eta_j(f, z) + \ln \delta(f, z)) + \varepsilon_j\right] \tag{10}$$

where $\varepsilon_j = \beta(z)\widetilde{\varepsilon}_j$ so $E(\varepsilon_j \mid r, z) = 0$, and f is now determined by

$$f = \arg\max \Psi\left(\widetilde{R}_{1f}, ...\widetilde{R}_{Jf}\right), \text{ where } \widetilde{R}_{jf} = R_j(f, y, r, z) + (M_j(f, z)/\beta(z))\varepsilon_j + \widetilde{e}_{jf} \tag{11}$$

We will want to estimate the Engel curve equations (10) for $j = 1, ..., J$. Equation (11) shows that f is an endogenous regressor in these equations, because f depends on both ε_j and $\widetilde{e}_{jf}$. As discussed in the main text, we do not try to empirically identify or estimate equation (11), because both the functions R_j and errors $\widetilde{e}_{1f}$ depend on u_j, and there may be important determinents of u_j (the direct utility or disutility from cooperation) that we cannot observe. However, we will require at least one instrument for f.

Another source of error in our model is that, in our data, y is a constructed variable (including imputations from home production), and so may suffer from measurement error. We will therefore require instruments for y. Our current collective household model is static. This is justified by a standard two stage budgeting (time separability) assumption, in which households first decide how much of their income and assets to save versus how much to spend in each time period, and then allocate their expenditures to the various goods they purchase. The total they spend in the time period is y, and the household's allocation of y to the goods they purchase is given by equation (6) in the main text. These means that variables associated with household income and wealth will correlate with y and so are potential instruments for y.

This time separability applies to the utility functions over goods, $U_j(q_j, g_j, z)$ for each member j, but need not apply to the utility or disutility associated with f, that is, $u_j(f, v, z)$. So at least some of these income and wealth variables could be components of v. Let $\widetilde{r}$ denote a vector of potential instruments for y. These are measures related to income or wealth that are not already included in v.

Assume there exists values v_0 and v_1 such that $u_j(f, v_0, z) \neq u_j(f, v_1, z)$ for some member j who's utility appears in Ψ. Then it follows from equation (11) that f varies with v, so v can serve as an instrument for f. Similarly, assume that $\ln y$ correlates with $\widetilde{r}$, which can serve as instruments for $\ln y$ (elements of v could also be instruments for y). Based on equation (10), we then have conditional moments

$$E\left[\left(\frac{w_j}{\eta_j(f,z)}-\gamma_j(z)-\beta(z)(\ln y+\ln\eta_j(f,z)+\ln\delta(f,z))\right)\mid\widetilde{r},v,z\right]=0 \tag{12}$$

Later in this Appendix we consider nonparametric identification of the functions in this expression based on these moments, but for now consider using these moments parametrically. If we parameterize each of the unknown functions using a parameter vector θ, then equation (12) implies unconditional moments

$$E\left[\left(\frac{w_j}{\eta_j(f,z,\theta)}-\gamma_j(z,\theta)-\beta(z,\theta)(\ln y+\ln\eta_j(f,z,\theta)+\ln\delta(f,z,\theta))\right)\phi(\widetilde{r},v,z)\right]=0 \tag{13}$$

for any suitably bounded functions $\phi(\widetilde{r},v,z)$. Our actual estimator will consist of parameterizing the unknown functions in this expression, choosing a set of functions $\phi(\widetilde{r},v,z)$, and estimating the parameters by GMM (the generalized method of moments) based on these moments. At the end of this Appendix we discuss choice of the ϕ functions.

Equation (13) can suffice for parametric identification and estimation, but is it still possible to nonparametrically identify the functions in this model in the presence of unobserved heterogeneity? The following Theorem shows that the answer is yes, if we make some additional assumptions. Theorem 1 shows these additional assumptions are sufficient for nonparametric identification of these functions, These additional assumptions, which are not required for parametric identification, are listed in Assumption A8.

ASSUMPTION A8. Add unobservable heterogeneity terms $\widetilde{\varepsilon}_j$ and $\widetilde{e}_{jf}$ to the model by replacing the function $\ln S_j(\pi_j,A_fp,z)$ with $\ln S_j(\pi_j,A_fp,z)-\widetilde{\varepsilon}_j$ and $u_j(f,v,z)$ with $u_j(f,v,z)+\widetilde{e}_{jf}$, for $j=1,..,J$. Assume f is determined by maximizing Ψ, where Ψ is linear, so $\Psi\left(\widetilde{R}_{1f},...\widetilde{R}_{Jf}\right)=\sum_{j=1}^{J}\widetilde{c}_j\widetilde{R}_{jf}$ for some constants $\widetilde{c}_1$,...,$\widetilde{c}_J$. Let $\widetilde{e}=\sum_{j=1}^{J}\widetilde{c}_j(\widetilde{e}_{j1}-\widetilde{e}_{j0})$. Define $\widetilde{y}(\widetilde{r},v,z)$ by $\ln\widetilde{y}(\widetilde{r},v,z)=E(\ln y\mid\widetilde{r},v,z)$. Assume the following: The function $\widetilde{y}(\widetilde{r},v,z)$ is differentiable in a scalar $\widetilde{r}$ with a nonzero derivative. The error $\widetilde{e}$ is independent of $y,\widetilde{r},v,z$ and $(\varepsilon_j,\widetilde{e})$ is independent of $\widetilde{r}$ conditional on (v,z). $E(\varepsilon_j\mid\widetilde{r},v,z)=0$. The functions $M_j(f,z)$ do not depend on f. There exist values v_1 and v_0 of v such that $\sum_{j=1}^{J}\widetilde{c}_ju_j(f,v_1,z)\neq\sum_{j=1}^{J}\widetilde{c}_ju_j(f,v_0,z)$.

THEOREM 1: Let Assumptions A1 to A8 hold. Then the functions $\eta_j(f,z)$, $\delta(f,z)$,$\gamma_j(z)$, and $\beta(z)$ are identified.

To prove Theorem 1, first observe that, with f binary, it follows from equation (11) that $f=1$ if $\sum_{j=1}^{J}\widetilde{c}_j[R_j(1,y,r,z)+(M_j(1,z)/\beta(z))\varepsilon_j+\widetilde{e}_{j1}]$ is greater than $\sum_{j=1}^{J}\widetilde{c}_j[R_j(0,y,r,z)+(M_j(0,z)/\beta(z))\varepsilon_j+\widetilde{e}_{j0}]$, where the function R_j is given by Lemma 6. Taking the difference in these expressions, and using the assumption that $M_j(f,z)$ doesn't depend on f, we get that $f=1$ if and only if

$$\begin{aligned}\sum_{j=1}^{J}\widetilde{c}_j[&(\ln\eta_j(1,z)+\ln\delta(1,z))M_j(z)+\mu_j(1,v,z)\\&-(\ln\eta_j(0,z)+\ln\delta(0,z))M_j(z)-\mu_j(0,v,z)]+\widetilde{e}\end{aligned}$$

is positive. This means that $f = \widetilde{f}(v, z, \widetilde{e})$ for some function $\widetilde{f}$. More precisely, f obeys a threshold crossing model where f is one if a function of v and z given by the above expression is greater than $-\widetilde{e}$, otherwise f is zero.

Now, again exploiting that f is binary,

$$E(w_j \mid \widetilde{r}, v, z, y) = E\left[W_j(f, z, y) + \beta(z) \ln \delta(f, z) \widetilde{\varepsilon}_j \mid \widetilde{r}, v, z, y\right]$$

$$= E[W_j(1, z, y) f + \beta(z) \ln \delta(1, z) f \widetilde{\varepsilon}_j + W_j(0, z, y)(1 - f) + \beta(z) \ln \delta(0, z)(1 - f) \widetilde{\varepsilon}_j \mid \widetilde{r}, v, z, y]$$

$$\begin{aligned} &= W_j(0, z, y) + \left[W_j(1, z, y) - W_j(0, z, y)\right] E(f \mid \widetilde{r}, v, z, y) \\ &+ \beta(z)\left[\ln \delta(1, z) - \ln \delta(0, z)\right] E(f \widetilde{\varepsilon}_j \mid \widetilde{r}, v, z, y). \end{aligned}$$

Next, observe that, since $W_j(f, z, y)$ is linear in $\ln y$, $E\left[W_j(0, z, y) \mid \widetilde{r}, v, z\right] = W_j(0, z, \widetilde{y})$ and $E\left[W_j(1, z, y) \mid \widetilde{r}, v, z\right] = W_j(1, z, \widetilde{y})$ where $\widetilde{y} = \widetilde{y}(\widetilde{r}, v, z)$. Averaging the above expression over y, and noting that $f = \widetilde{f}(v, z, \widetilde{e}_1)$, we get

$$\begin{aligned} E(w_j \mid \widetilde{r}, v, z) &= W_j(0, z, \widetilde{y}) + \left[W_j(1, z, \widetilde{y}) - W_j(0, z, \widetilde{y})\right] E(f \mid \widetilde{r}, v, z) \\ &+ \beta(z)\left[\ln \delta(1, z) - \ln \delta(0, z)\right] E(f \widetilde{\varepsilon}_j \mid \widetilde{r}, v, z). \end{aligned}$$

and by the conditional independence assumptions regarding $\widetilde{\varepsilon}_j$ and $\widetilde{e}_1$,

$$\begin{aligned} E(w_j \mid \widetilde{r}, v, z) &= W_j(0, z, \widetilde{y}) + \left[W_j(1, z, \widetilde{y}) - W_j(0, z, \widetilde{y})\right] E(f \mid v, z) \\ &+ \beta(z)\left[\ln \delta(1, z) - \ln \delta(0, z)\right] E(f \widetilde{\varepsilon}_j \mid v, z). \end{aligned}$$

Now the functions $E(w_j \mid \widetilde{r}, v, z)$ and $\widetilde{y}(\widetilde{r}, v, z)$ (the latter defined by $\ln \widetilde{y}(\widetilde{r}, v, z) = E(\ln y \mid \widetilde{r}, v, z)$) are both identified from data (and could, e.g., be consistently estimated by nonparametric regressions. So the derivatives of these expressions with respect to $\widetilde{r}$ are identified. This means that the following expression is identified.

$$\frac{\partial E(w_j \mid \widetilde{r}, v, z)}{\partial \ln \widetilde{r}} \Big/ \frac{\partial \ln \widetilde{y}(\widetilde{r}, v, z)}{\partial \ln \widetilde{r}} = \frac{\partial W_j(0, z, \widetilde{y})}{\partial \ln \widetilde{y}} + \frac{\partial\left[W_j(1, z, \widetilde{y}) - W_j(0, z, \widetilde{y})\right]}{\partial \ln \widetilde{y}} E(f \mid v, z) \tag{14}$$

Taking the difference between the above expression evaluated at $v = v_1$ and at $v = v_0$ then gives (and so identifies)

$$\frac{\partial\left[W_j(1, z, \widetilde{y}) - W_j(0, z, \widetilde{y})\right]}{\partial \ln \widetilde{y}} \left[E(f \mid v_1, z) - E(f \mid v_0, z)\right]$$

and, since $E(f \mid v, z)$ is also identified, this identifies $\partial\left[W_j(1, z, \widetilde{y}) - W_j(0, z, \widetilde{y})\right] / \partial \ln \widetilde{y}$. We can then solve equation (14) for $\partial W_j(0, z, \widetilde{y}) / \partial \ln \widetilde{y}$ where all the terms defining this derivative are identified. Taken together, the last two steps identify $\partial W_j(f, z, \widetilde{y}) / \partial \ln \widetilde{y}$ for $f = 0$ and for $f = 1$.

Given these identified functions and derivatives, we may then duplicate the proof of Lemma 5, (replacing y with $\widetilde{y}$, to show that the functions $\beta(z)$, $\eta_j(f, z)$, $\gamma_j(z)$, and $\delta(f, z)$ are identified.

2 Instrument Validity

To more formally define conditions under which village-level average f is a valid instrument, assume that the household h random utility parameters $\widetilde{e}_{1fh}$ and $\widetilde{\varepsilon}_{jh}$ defined in the Appendix are independent across households. Let $\overline{f}_h$ equal the expected value of f_h conditional on being a household other than h in the village. Then $\overline{f}_h$ is the probability that a randomly chosen household in the village, other than household h, cooperates. Assume that we include $\overline{f}_h$ in the function R_j (equals $U_j + u_j$, formally defined in the Appendix). Taking the conditional mean of Equation (11) across households other than household h in the village then shows that $\overline{f}_h$ equals a function of the joint distribution of $y_{h'}$, $\boldsymbol{r}_{h'}$, $\boldsymbol{z}_{h'}$, $\widetilde{e}_{1fh'}$ and $\widetilde{\varepsilon}_{1h'}$ across all households h' other than h in the village. It follows that $\overline{f}_h$ is a relevent instrument in that it affects the choice of f (by being in R_j) and that it is a valid instrument in the quantity demand equations because $\overline{f}_h$ is independent of household h's specific value of $\widetilde{\varepsilon}_{jh}$ and hence of ε_{jh}.