Identification of Random Resource Shares in Collective Households Without Preference Similarity Restrictions

Geoffrey R. Dunbar, Arthur Lewbel, and Krishna Pendakur
Bank of Canada, Boston College, and Simon Fraser University

Original Oct 2015, Revised June 2019

Abstract

Resource shares, defined as the fraction of total household spending going to each person in a household, are important for assessing individual material well-being, inequality and poverty. They are difficult to identify because consumption is measured typically at the household level, and many goods are jointly consumed, so that individual level consumption in multi-person households is not directly observed. We consider random resource shares, which vary across observationally identical households. We provide theorems that identify the distribution of random resource shares across households, including children’s shares. We also provide a new method of identifying the level of fixed or random resource shares that does not require previously needed preference similarity restrictions or marriage market assumptions. Our results can be applied to data with or without price variation. We apply our results to households in Malawi, estimating the distributions of child and of female poverty across households.

JEL codes: D13, D11, D12, C31, I32. Keywords: Collective Household Model, Cost of Children, Bargaining Power, Sharing Rule, Demand Systems, Engel Curves

1 Introduction

Expenditure surveys generally collect consumption data at the level of households. Standard poverty and welfare measurements based on such data are also typically calculated at the household level. But well-being and utility apply to individuals, not households. When household resources are distributed unequally across household members, official household level measures of income or consumption can seriously underestimate the prevalence of
poverty and inequality in a country. For example, using the types of methods that we will extend here, Dunbar, Lewbel and Pendakur (2013, hereafter denoted DLP), find poverty rates for children in Malawi that are much higher than those of men. Another example is Calvi (2017), who finds that the poverty rate among older married women in India increases with age, primarily because their share of household resources declines with age.

The resource share of a given household member is defined as the fraction of total household consumption expenditures that are spent on goods and services that the given household member consumes. Resource shares are in general difficult to identify, particularly children’s resource shares, because consumption is measured typically at the household level, and many goods are jointly consumed and/or shareable.

In this paper we provide two new theorems on the identification of resource shares within households, and provide associated estimators. Our first theorem allows for unobserved heterogeneity in resource allocations across households. By treating resource shares as random variables, our first theorem shows identification of the distribution (around an unknown level) of resource shares across households. Our second theorem shows identification of the level of resource shares. Unlike existing related theorems, this second theorem does not require assumptions regarding similarity of tastes across households of different composition or across people within a household. This second theorem instead uses a new way to exploit variation in so-called distribution factors, that is, variables that affect the allocation of resources within households but do not affect the tastes of individuals. Taken together, our theorems completely identify the joint distribution of resource shares across households.

We apply our results to household survey data from Malawi. We find that allowing for unobserved heterogeneity in resource allocations across households matters for welfare analysis in two important ways. First, we find that unobserved heterogeneity accounts for about the same amount of variation in resource shares as does observed heterogeneity. Second, we find that measured poverty rates are higher when allowing for random resource shares rather than constraining them to be fixed.

In addition to identifying the level and distribution of resource shares, we also use our results to test some of the behavioural restrictions that previous work required for identification. We find evidence supporting the similarity across people (SAP) restriction that was proposed by DLP for identification.

We start with the Pareto efficient type of collective household model of Becker (1965, 1981) and Chiappori (1988, 1992). The earlier literature on these models, including Browning, Bourguignon, Chiappori, and Lechene (1994), Browning and Chiappori (1998), Vermeulen (2002), and Chiappori and Ekeland (2009), showed that, even if one knew the entire demand vector-function of a household (that is, how much the household buys of every good
and service as a function of prices, income, and other observed covariates), one still cannot identify the level of each household member’s resource. However, this earlier work also shows that one can generally identify how these resource shares would change in response to a change in observed covariates such as distribution factors. Other papers that make use of this result include Bourguignon and Chiappori (1994), Chiappori, Fortin and Lacroix (2002), and Blundell, Chiappori and Meghir (2005). Most of this earlier work also constrains goods to be either purely private or purely public within a household, whereas we work with a more general model based on Browning, Chiappori and Lewbel (2013) and DLP, that allows some or all goods to be partly or fully shared.

Our first theorem extends this earlier literature by showing that, though the level of resource shares is not identified without additional assumptions, one can identify the distribution of random variation in these shares (around the unknown levels) across households. Two prior results exist on identification of the distribution of random resource share, both of which are much more restrictive than ours. One, by Matzkin and Perez-Estrada (2011), is based on semiparametric restrictions of functional forms and focuses on variation in a single good. The other, by Chiappori and Kim (2017) is closer to ours but unlike our Theorem 1, requires observed price variation, deals only with two member households, and requires every good to be either purely private or purely public.

As noted above, from just observing household demand functions, one cannot identify the absolute levels of resource shares; only variations in resource shares around those levels are identified. Some interesting policy questions can be addressed without identifying levels, such as identifying whether a policy that changes a distribution factor, like women’s access to contraception, actually increases women’s share of resources within the household. However, other fundamental policy questions, such as identifying the prevalence of women’s or children’s poverty, requires identifying resource share levels. Our Theorem 2 provides minimal additional assumptions needed to identify the levels of resource shares.

A number of previous approaches exist to address the fundamental nonidentification of resource share levels just from household demand data. One direct approach, taken e.g. by Menon, Perali and Pendakur (2012) and Cherchye, De Rock and Vermeulen (2012), is to collect as much detailed data as possible on the separate consumption of each household member, rather than just observing household-level consumption. To the extent that such data can be collected, resource shares may be observed directly, by taking each individual’s observed consumption as a fraction of the total. However, this method requires detailed and difficult data collection, and is likely to suffer from considerable measurement errors, particularly in the allocation of public and shared goods to individual household members.

A second approach is taken by Cherchye, De Rock and Vermeulen (2011). While the
levels of resource shares cannot be identified without additional assumptions, these authors show that it is possible to obtain bounds on resource shares, using revealed preference inequalities. Cherchye, De Rock, Lewbel, and Vermeulen (2015) considerably tighten these bounds by combining Slutsky symmetry type restrictions with revealed preference inequalities. Bounds can be further tightened, sometimes leading to point identification of shares, by further combining revealed preference inequalities with assumed restrictions regarding marriage markets. See, e.g., Cherchye, De Rock, Demuynck, and Vermeulen (2014).

A third method is to completely identify the level of resource shares from household level data by imposing additional restrictions either on preferences, or on the household’s allocation process, or both. Papers that use these methods include Lewbel (2003), Lewbel and Pendakur (2008), Couprie, Peluso and Trannoy (2010), Bargain and Donni (2009, 2012), Lise and Seitz (2011), Browning, Chiappori and Lewbel (2013), and DLP. These methods all employ either strong restrictions on the functional forms of preferences, or assumptions regarding similarity of preferences across households of different composition, or similarity of tastes across people within a household. For example, Lise and Seitz (2011) impose a particular functional form on utility functions, Browning, Chiappori and Lewbel (2013) assume individual’s indifference curves over goods do not change when they marry, and DLP assume tastes for an assignable good like clothing are similar either across people within a household, or across people of the same type across households. Most of these papers that impose preference restrictions on individuals do not require or exploit distribution factors.

Both Lewbel and Pendakur (2008) and DLP make an additional assumption that resource shares do not depend on the total expenditures of the household (though they can still depend on the total income or wealth of the household). This assumption allows them to use data without price variation, i.e., to work with Engel curves instead of Marshallian demand functions. This considerably reduces data requirements (relative to all other papers that identify resource share levels) by, i) greatly simplifying the demand functions that need to be estimated, ii) eliminating the need to collect price data, and, iii) allowing estimation using data from a single cross section survey. Evidence supporting this assumption empirically can be found in intra-household data collection papers like Menon, Perali and Pendakur (2012) and in the bounds estimates of Cherchye, De Rock, Lewbel, and Vermeulen (2015).

Our second theorem provides a new set of minimal conditions that suffice for identification of the level of resource shares. This theorem shows that if one maintains just the assumption that resource shares not depend on total expenditures, and observes some assignable goods and some distribution factors, then that alone can be enough to identify the level of resource shares. No similarity restrictions on tastes like those discussed above are needed. Moreover, like Lewbel and Pendakur (2008) and DLP, these conditions permit identification without
requiring any price variation. Because this new identification result does not impose these preference similarity restrictions, it allows us to empirically test those constraints.

After establishing identification, we show how these models can be estimated and tested. A general semiparametric method of estimating individual’s demand functions within a household, along with the distribution of resource shares, is given in the appendix. In the main text we provide a simpler estimator that combines a commonly used functional form for collective household model demands with nonparametric estimation of the random variation in resource shares.

The next sections set up the model and provide our identification theorems. We then describe our proposed estimators. This is followed by empirical results from applying our model to a survey of households in Malawi. We use the results to test the types of preference restrictions that earlier identification results required. We also report estimates of the distribution of child poverty (and female poverty) across households which account for random variation in resource shares.

2 Resource Shares

A key component of collective household models, going back to the earliest frameworks of Becker (1965, 1981) and Chiappori (1988, 1992), are resource shares. Resource shares are defined as the fraction of a household’s total resources (spent on consumption goods) that are allocated to each household member. Resource shares, which are closely related to Pareto weights, are often interpreted as measures of the bargaining power of each household member, however, they are also determined by altruism, particularly the shares claimed by children. Our model starts with the Pareto efficient collective household model of Browning, Chiappori and Lewbel (2013), hereafter BCL.

2.1 BCL and DLP

Unlike earlier collective household models, BCL does not require goods to be purely public or purely private, but instead permits goods to be partly shared, using what is called a consumption technology function. Like earlier results in this literature, BCL show that as a result of the Pareto efficiency of the household’s resource allocation process, maximizing the household’s objective function is observationally equivalent to a decentralised allocation. In this allocation, each household member demands a vector of consumption quantities given their preferences and a personal budget constraint, and the household purchases the sum of these demanded quantities (adjusted to account for the extent to which goods are shared or
consumed jointly).

Given this decentralization, we can conceptualise the household’s behavior as creating budget constraints for its members. These budget constraints, combined with the preferences of individual household members, are sufficient to conduct ordinal welfare analysis. Each person’s budget constraint is characterised by a shadow budget and a shadow price vector. These are “shadow” budgets and prices because they govern each agent’s consumption demands but are not observed and do not equal the observed household budget or market prices.

Shadow budgets must add up to the full household budget. Each person’s share of the household budget is called their resource share. Resource shares may be unequal across household members, due to individual variation in bargaining power and altruism within the household. BCL show that resource shares have a one-to-one correspondence with Pareto weights on individual utilities in the overall household maximization problem. Resource shares may depend on distribution factors, defined as variables that affect bargaining power or claims on resources within the household, but do not affect preferences for goods and services.

In the BCL model, shadow prices for goods must be the same for all household members. (If they were not the same, then there would be gains from trade across household members, a violation of the assumption of efficiency.) Shadow prices are determined by the consumption technology of the household, and are less than or equal to market prices. The more a good is shared, the lower is its shadow price. For goods that are not shared, the shadow price equals the market price. In their empirical application, BCL impose the restriction that the shadow price vector is a linear transformation of market prices. We do the same here. This ensures that a linear household budget constraint given market prices implies a linear shadow budget constraint for each household member.

BCL fully identify both the consumption technology function and the resource shares, so that they can completely characterise the shadow budget constraints faced by each household member. They accomplish this by imposing strong restrictions on the similarity of preferences between individuals living alone vs living together, requiring data on both the demands of singles and of couples, and by requiring substantial observed relative price variation. More recent work focuses on relaxing these assumptions and data requirements, and as a result only on identifying some particular features of interest, such as resource shares. Our paper falls in this category.

Lewbel and Pendakur (2008) and DLP modified BCL to permit identification of resource shares from data that do not contain price variation (Engel curve type data), by placing restrictions on how prices and the consumption technology function interact, and by imposing
the constraint that resource shares not vary with total expenditures. Both provide theoretical and empirical evidence supporting this identifying assumption. DLP also substantially relaxed the BCL restriction limiting differences in preferences, and further relaxed data and estimation requirements by only needing to observe and estimate household Engel curves on one private, assignable good for each household member. Note that private goods are defined as goods that are not shared, and an assignable good is one where we can observe which household member consumes the good.

Unlike BCL, DLP permits identification of resource shares of both adults and of children, where children are treated as having their own utility functions and welfare. This is in contrast to most of the empirical collective household literature, where expenditures on children are modeled just as public goods in the adults’ utility functions. The identification of children’s resource shares is necessary to answer questions regarding the welfare of children in the household, separate from the welfare of the parents.

The present paper, like DLP, is a model with both adult’s and children’s resource shares, in a data environment based on observable assignable goods that does not require price variation. As discussed in the introduction, the present paper generalises DLP in two important ways. One is the identification of resource shares without imposing the BCL or DLP restrictions on similarity of preferences across households of different composition, or across people within a household. The other is by allowing resource shares to vary randomly across households, equivalent to allowing for the existence of unobserved distribution factors, or unobserved heterogeneity in bargaining power or altruism.

2.2 Children and Random Resource Shares

Let $d$ denote a vector of distribution factors. Distribution factors are important variables in the collective household literature for three reasons. First, distribution factors are closely related to individual’s relative bargaining power within households, and so are important components of marriage market models, and models of women’s empowerment (particularly in poor and developing countries). As a result, distribution factors arise in other literatures associated with household formation, stability, and function. Second, some types of distribution factors are policy variables, providing governments with the ability to affect the within-household distribution of resources. Examples can include inheritance and divorce laws, access to credit or banking for women, and local availability of education or medical care. Third, in efficient collective households the response of resource shares to a change in distribution factors can generally be identified just from household-level demand behaviour, given many observations of the same or fully comparable household(s) in different price and
income regimes, without additional assumptions.

Our model allows for random variation in resource shares across households. Equivalently, we allow for the presence of unobserved distribution factors, which we will denote $v$. This reflects that fact that the relative bargaining power of individual household members, or the choice of social welfare function for the household, will in general vary across households for unobserved reasons. So, e.g., two households that otherwise appear identical could have different values of resource shares, due to unobserved differences in household member’s personality traits, beauty, intelligence, altruism, etc. Both $d$ and $v$ are defined to be variables that affect resource shares, but do not affect preferences.

Let the household be comprised of $J$ individuals indexed by $j$. Our typical case will be $J = 3$, consisting of a man, a woman, and children.\footnote{Due to data limitations in our empirical analysis, we treat all of the children as having a single utility function, or equivalently, a unitary joint social welfare function. Other datasets could potentially allow researchers to relax this restriction, by observing some component of consumption separately for each child.} Let $x^*$ be the total expenditures of the household, i.e., the household’s total budget.

Let $z$ denote a vector of observable attributes of households and their members like age, education, and number of children. Household attributes $z$ may affect the preferences of each household member and may also affect the household’s bargaining process or social welfare function, and as a result may directly affect resource shares.

Let scalars $y_1,...,y_J$ denote quantities of private, assignable goods. Specifically, each $y_j$ is the quantity of some good $j$ that is only consumed by household member $j$. What makes each $y_j$ assignable is that each appears in just one (known) person’s utility function, and what makes each $y_j$ private is that it is not shared or consumed jointly. So, e.g., in a household with a woman, a man, and a child, the private assignable goods could be women’s clothes, men’s clothes, and children’s clothes.

Let $p$ denote the vector of prices of all the goods and services that the household buys, sorted so that the prices of assignable goods are the first $J$ elements of $p$. Thus, $p$ includes the prices $p_1,...,p_J$ of the private assignable goods, as well as the prices of all other goods the household buys. Let $\eta_j(p,x^*,z,d,v)$ denote the resource share of member $j$, that is, $\eta_j$ is the fraction of total household resources $x^*$ that are allocated to member $j$.

Adapting BCL and DLP, in the Appendix we show that the household demand functions of private assignable goods are given by

$$y_j = h_j(\eta_j(p,x^*,z,d,v)x^*,p,z) \text{ for } j = 1,...,J. \quad (1)$$

The interpretation of this equation is that the resources allocated to member $j$ are $\eta_j x^*$ (the share $\eta_j$ of total household budget $x^*$) and the function $h_j$ is that member’s Marshallian...
demand function for this good. Since this good is private and assignable, the household’s demand for the good equals just that member’s own demand for the good. It is important to note that only private assignable goods have the simple form given by equation (1). The demand functions for other goods are much more complicated, as in BCL.

We will assume that resource shares \( \eta_j \) are independent of the household budget \( x^* \). Lise and Seitz (2011), Lewbel and Pendakur (2008), Bargain and Donni (2009, 2012) and DLP all use this restriction in their identification results, and supply some theoretical arguments for it. Cherchye, De Rock, Lewbel and Vermeulen (2015) and Menon, Perali and Pendakur (2012) provide empirical support for this restriction using American and Italian data, respectively. Also reducing the restrictiveness of this assumption is that resource shares are only assumed to be independent of \( x^* \) after conditioning on covariates such as \( z \) and \( d \), both of which could include variables closely related to \( x^* \), such as wealth, income (which equals \( x^* \) plus savings), education, wages, etc.

Our model allows for a certain type of measurement error in the household budget. Let \( x \) denote the observed household budget, which is related to the true household budget \( x^* \) by

\[
x^* = e x,
\]

where \( e \) is measurement error, which we assume has mean one. We assume this error is of the Berkson (1950) type, meaning that it is distributed independently of \( x \). We assume measurement error takes a multiplicative form because larger levels of expenditures are likelier to be associated with larger errors. Our main reason for assuming Berkson errors is pragmatic; with our particular model and identifying assumptions, we can allow for this type of error essentially for free, i.e., with no additional assumptions and no added complexity in estimation. It may be possible to also allow for ordinary (classical) measurement error using a control function approach, though we do not pursue that possibility here.

The assumption that resource shares are independent of the household budget means we may drop \( x^* \) from the \( \eta_j \) function, and our total expenditure measurement error model means we can replace \( x^* \) with \( e x \) in equation (1). The result is

\[
y_j = h_j (\eta_j(p, z, d, v) e x, p, z) \quad \text{for } j = 1, \ldots, J.
\]

We assume that we observe data \( \{y_1, \ldots, y_J, p, z, d, x\} \) for a sample of households. However, if all households in the sample face the same price vector, as in a cross section of Engel

\footnote{Note that this \( h_j \) function may depend on \( z \) not just because tastes may vary with \( z \), but also because households can vary in the extent to which they share goods other than the private assignable goods, and so the \( h_j \) functions include accounting for the resulting variation in shadow prices relative to market prices \( p \).}
curve data, then it is unnecessary to observe $p$, and all of our results will go through just conditioning on that period’s prices.

The random unobservables in our model consist of the measurement error $e$ and the unobserved distribution factors $v$. Since the resource shares $\eta_1, \ldots, \eta_J$ depend on the unobserved random vector $v$, these resource shares are themselves random variables. Our goal is to nonparametrically identify the conditional distribution of $\eta_1, \ldots, \eta_J$, conditioning on $z, d$ and, if price variation is observed, $p$. The identification is nonparametric in that the theory assumes the functional forms of $h_1, \ldots, h_J$ are unknown, and functional forms of how $\eta_1, \ldots, \eta_J$ depend on $p, z, d$ and $v$ are also unknown.

Our identification argument proceeds in two steps. First we show the distribution of $\eta_1, \ldots, \eta_J$ is identified around unknown conditional mean functions. The distribution of the measurement error $e$ is also nonparametrically identified. Then we provide sufficient conditions for identifying these conditional mean functions, without imposing the DLP or BCL type preference similarity restrictions, by exploiting variation in $d$. The advantage of separating these two results is that identification of the conditional mean functions requires more assumptions. Also, behavioural restrictions like those in DLP or BCL could be used instead of, or in addition to, the method we provide here for identifying the conditional means of the resource shares.

In an appendix we describe a general estimator based on these identification results. In the main text we provide an estimator that is easier to implement, but is based on a particular popular functional form for the demand functions $h_j$. Both estimators are nonparametric regarding the the distribution of resource shares $\eta_1, \ldots, \eta_J$ and of the measurement error $e$.

3 Nonparametric Identification of Resource Shares and Their Distribution

In this section we first give some intuition for how we use observed distribution factors to identify the levels of deterministic resource shares, providing an alternative to the restrictions on preferences that were required by BCL, DLP, and others to identify levels. We then allow resource shares to vary randomly across households (corresponding to the presence of unobserved distribution factors) and give some intuition for how we can identify the joint distribution of these resource shares. This is then followed by our formal list of assumptions and identification theorems.
3.1 Identification Intuition With Fixed Resource Shares

To simplify notation in this subsection, assume we have a sample with no variation in \( p, z, \) or \( d \). Assume also for now that there is no measurement error, so \( e = 1 \), and no unobserved distribution factors \( v \) (and therefore no random variation in resource shares). Equation (2) then simplifies to

\[
y_j = G_j(x) = h_j(\eta_j x) \quad \text{for } j = 1, ..., J.
\]

where \( G_j(x) \) is an observable household demand function for person \( j \)’s private assignable good. The goal is then identification of \( \eta_j \), which is now just a constant for each household member \( j \).

3.2 BCL and DLP Identification

In this greatly simplified version of the model, the BCL identification would correspond to identifying the \( h_j \) functions by looking at singles (people living alone, for whom \( \eta_j = 1 \)). Then, with \( h_j \) known, one could solve equation (3) for \( \eta_j \). This requires assuming that individual’s tastes for goods, and therefore the \( h_j \) functions (apart from the consumption technology) do not change when people form collective households (or, in particular, when they marry).

Note we are just providing intuition here; the actual identification in BCL is more difficult because (among other restrictions) they do not require the existence of private assignable goods, they do not impose the constraint that \( \eta_j \) not depend on \( x \), and they explicitly account for the consumption technology (i.e. sharing) of other goods. Other papers like Lewbel and Pendakur (2008), Donni et. al. (2012) and Pendakur (2018) similarly make use of singles to identify \( h_j \) and thereby recover \( \eta_j \).

DLP employ (two) different identification strategies that do not depend on singles. One of DLP’s approaches is to instead suppose that \( h_j \) is similar across people (SAP) within the household. An extreme case would be if the \( h_j \) function was the same across all the \( j = 1, ..., J \) people in the household, so we’d have \( y_j = h(\eta_j x) \) instead of \( y_j = h_j(\eta_j x) \) for some function \( h \). Here, there are \( J - 1 \) unknown scalars \( \eta_j \) and one unknown function \( h \) to identify from \( J \) observed functions \( G_j \), so we meet the order condition for identification.

In particular, variation in \( x \) allows us to identify \( \eta_j \). To illustrate, suppose \( h \) was linear so \( h(\eta_j x) = b\eta_j x + a \) for unknown coefficients \( b \) and \( a \). Then the observable demand functions (for these private assignable goods) are \( G_j(x) = c_j x + a \) where \( c_j = b \eta_j \). The coefficient of \( x \) identifies \( c_j \). Using the fact that resource shares \( \eta_j \) sum to one, we can identify \( b \) by \( \sum_{j=1}^J c_j = \sum_{j=1}^J b \eta_j = b \). Then, once we know \( b \) and each \( c_j \), we identify each \( \eta_j \) by \( \eta_j = c_j / b \). Essentially, we identify the \( J \) coefficients \( c_1, ..., c_J \), and from those we obtain the \( J \) parameters...
b, \eta_1,...,\eta_{J-1}. The last parameter \eta_J is obtained from the constraint that shares sum to one.

For nonlinear h, we can do a similar construction with derivatives (denoted by apostrophes). For example, if \( G_j(x) = h(\eta_j x) \) then \( G'_j(x) = \eta_j h'(\eta_j x) \). So \( G'_j(0) = \eta_j h'(0) \) and we can treat \( h'(0) \) like b above. Again we are just providing intuition; the actual identification in DLP is more complicated than this illustration. For example, DLP allows for \( h_j \) functions that are not identical across people but merely similar (i.e., only having some coefficients in common). DLP also consider cases where, instead of being similar across household members, demands are similar across types (SAT), e.g., women having similar demands regardless of whether the household has one, two, or three children.

3.3 Identification Using Observed Distribution Factors

Both BCL and DLP (and other related results in the literature) depend heavily on some preference restrictions, like tastes not changing when forming couples, or similarity of tastes across household members. But, unlike the earlier literature that only identified changes but not levels in resource shares, the newer identification results like BCL and DLP do not require distribution factors. Our new method for identifying levels of resource shares, given as Theorem 1 in the next section, does not impose the types of preference restrictions needed by results like BCL and DLP. Instead our new method exploits observable distribution factors while maintaining the assumption that resource shares not depend on \( x \).

For intuition on how Theorem 1 will work, continue to assume a sample with no variation in \( p \) or \( z \), no measurement error, and no unobserved distribution factors \( v \), but now assume we see some variation in distribution factors \( d \), so the assignable good demand functions are now

\[
y_j = G_j(d, x) = h_j(\eta_j x) \quad \text{for} \quad j = 1, ..., J.
\]

where \( G_j(d, x) \) is an observable demand function. The goal is identification of \( \eta_{dj} \), which now is just a different constant for each value that \( d \) and \( j \) can take on.

To keep things simple, suppose \( d \) is a discrete scalar that just takes the values 1, 2, ..., \( D \) for some integer \( D \). Again, just for intuition, consider the case where these \( h_j \) demand functions are linear, but we no longer impose the DLP similarity across people restriction, so \( h_j(\eta_{dj} x) = b_j \eta_{dj} x + a_j \) for unknown coefficients \( b_j \) and \( a_j \). Then the observable household demand functions (for these private assignable goods) are \( G_j(d, x) = c_{dj} x + a_j \) where \( c_{dj} = b_j \eta_{dj} \). For each value \( d \) can take on, the coefficient of \( x \) in the \( j \)'th equation identifies the coefficient \( c_{dj} \). We can identify \( DJ \) coefficients, i.e., \( c_{dj} \) for each value of \( d \) and \( j \). The unknown parameters are now \( b_1,...,b_J \), and, for each value of \( d \), \( \eta_{d1},...,\eta_{d,J-1} \) (again using that \( \eta_{dJ} = 1 - \sum_{j=1}^{J-1} \eta_{dj} \)). This gives a total of \( J + (J - 1)D \) unknowns. So if \( D \geq J \), meaning
that we have at least as many different values of the distribution factors as we have household members, then we have enough equations to identify all the resource shares $\eta_{dj}$.

The above shows we have at least as many equations as unknowns. This is a necessary but not sufficient condition for identification, so the identification proof also entails showing that a corresponding rank condition holds. For example, consider the case where $D = J = 2$. Then the equations $c_{dj} = b_j \eta_{dj}$ and $\eta_{d2} = 1 - \eta_{d1}$ together imply $c_{11}/c_{21} = \eta_{11}/\eta_{21}$ and $c_{12}/c_{22} = (1 - \eta_{11})/(1 - \eta_{21})$, which are two equations that, given some mild inequalities, we can readily solve for the two unknowns $\eta_{11}$ and $\eta_{21}$.

Our actual identification theorem extends this idea to nonlinear demand functions, by doing a similar analysis based on derivatives of $G_j(d,x)$ with respect to $x$ in place of linear coefficients of $x$. Finally, in all of these results, everything we call constants above are, in the real models, functions of $z$, and also of $p$ if the data contain price variation.

The main result here is that, instead of requiring preference similarity restrictions, all we need is an observable distribution factor vector $d$ that can take on as many values as we have family member types. For example, if $J = 3$ (man, woman, and children), then two binary distribution factors are enough.

### 3.4 Identification Intuition For Random Resource Shares

Now consider identification in the presence of unobserved distribution factors $v$, or equivalently, identification of the distribution of random resource shares. Again simplifying notation as much as possible for intuition, assume we have a sample with no variation in $p$, $z$, or $d$, and no measurement error, so $e = 1$. The private assignable demand functions are then

$$y_j = h_j(\eta_j(v)x) \quad \text{for} \quad j = 1, \ldots, J. \tag{5}$$

Here $v$ is a realization of the random, unobserved resource share vector $V$, and $y_j$ is the corresponding realization of a random quantity demand $Y_j$. For simplicity just write $\eta_j(V)$ as $\eta_j$ remembering now that $\eta_j$ is a random variable, not a constant. From data we can identify $F_Y(y_1, \ldots, y_J \mid x)$, that is, the joint distribution of $Y_1, \ldots, Y_J$, conditional on $x$. The goal is to identify the distribution of the random vector $\eta_1, \ldots, \eta_{J-1}$. Again since resource shares sum to one, if that distribution is known then we also know the distribution of the last resource share $\eta_J$.

What we show in our theorems is that knowledge of $F_Y(y_1, \ldots, y_J \mid x)$ is sufficient to back out the joint distribution of $\eta_1, \ldots, \eta_{J-1}$ up to an unknown location for each $\eta_j$. For example, just from $F_Y$ we could not determine the mean of each $\eta_j$, but we can determine its distribution around that unknown mean. The key element to showing this is that the
unobserved $V$ act on observed demands only through the resource shares $\eta_j$. If the functions $h_j$ were known, we could invert them to get $h_j^{−1}(Y_j) = \eta_j$, and the distribution of $\eta_j$ would be identified from the distribution of $h_j^{−1}(Y_j)$.

The key to our identification theorem is showing that the $h_j^{−1}$ function can be identified up to an unknown location, thereby letting us identify the distribution of the resource shares up to an unknown location for each share. We can then combine this identification with the results from the previous subsection (or with DLP or BCL), to get the unknown locations for each $\eta_j$, and thereby identify their entire joint distribution.

To see why it is possible to obtain the $h_j^{−1}$ function up to an arbitrary location, consider the probability that $Y_j \leq y$, conditional on $x$, for any given values of $y$ and $x$. That probability is something we can estimate from data, and by the model that probability equals the probability that $h_j(\eta_j x) \leq y$. For any change in $y$, we can calculate the corresponding change in $x$ that leaves this probability unchanged, and hence leaves the value of $h_j(\eta_j x)$ unchanged. This allows us to trace out how the function $h_j(\eta_j x)$ changes with $x$, identifying $h_j^{−1}$ function up to an arbitrary location. As in the previous subsection, in our real theorems we need to include $z$, $d$ (if any) and $p$ (if the data contain price variation) in the derivations.

Theorem 1 later formalizes this identification around unknown location, and Theorem 2 then adds the earlier described identification of location based on distribution factors. Using a very different proof technique, Chiappori and Kim (2017) also obtain an identification result for random resource shares around unknown locations, but their result requires substantial price variation, deals only with two member households, and has no measurement error $e$. So, e.g., their results could not be applied either to Engel curve data, or to a model with children.

### 3.5 Some Details

Here we give a few more details regarding distribution identification. Formal theorems and derivations are in the next subsection and the Appendix.

For exposition, focus on a single $j$, and continue to drop $p$, $d$, $z$, and $e$. Let $F_{Y_j}(y_j \mid x)$ be the distribution function of $Y_j$ given $x$ that we can identify and estimate from data, and let $F_{\eta_j}(\eta_j)$ denote the unknown distribution of the random resource share $\eta_j$. Assume the function $h_j$ is invertible. For any given value $y_j$ and $x$ we have that

$$F_{Y_j}(y_j \mid x) = \Pr(Y_j \leq y_j \mid x) = \Pr(h_j(\eta_j x) \leq y_j \mid x) = \Pr(\eta_j x \leq h_j^{−1}(y_j) \mid x) = \Pr\left(\eta_j \leq \frac{h_j^{−1}(y_j)}{x}\right) = F_{\eta_j}\left(\frac{h_j^{−1}(y_j)}{x}\right)$$

14
where the last equality uses the assumption that the distribution of $V$ and hence $\eta_j$ is independent of $x$. Let $\gamma = h_j^{-1}(y_j)/x$. We want to use this expression to learn about the $h_j$ function, but we don’t know $F_{\eta_j}$. However, we can use the above to get an expression only involving $h_j$. Consider the ratio of derivatives

$$\frac{-\partial F_{Y_j}(y_j | x) / \partial y_j}{\partial F_{Y_j}(y_j | x) / \partial \ln x} = \frac{-\partial F_{\eta_j}(h_j^{-1}(y_j)/x) / \partial y_j}{x \partial F_{\eta_j}(h_j^{-1}(y_j)/x) / \partial x} = \frac{-F_{\eta_j}'(h_j^{-1}(y_j)/x) h_j^{-1}(y_j)}{x F_{\eta_j}'(h_j^{-1}(y_j)/x) h_j^{-1}(y_j)/x^2} = h_j^{-1}(y_j) = \frac{\partial \ln h_j^{-1}(y_j)}{\partial y_j}$$

where the apostrophe denotes differentiation. Note the second equality applies the chain rule to both the numerator and denominator in the ratio of derivatives. The function $F_{\eta_j}$ drops out, leaving the log derivative of $h_j$ on the right hand side. The left side of this equation is something identified (that we can estimate from data), so this log derivative of $h_j^{-1}$ is identified. Integrating this expression with respect to $y_j$ then identifies the function $h_j$ itself, up to an unknown constant of integration. This unknown constant of integration is the unknown location or mean term discussed earlier.

In our data we have $J$ quantity demands $y_1, ..., y_J$ to work with, but we only needed to identify the distribution of $J - 1$ resource shares $\eta_1, ..., \eta_{J-1}$, because the last one $\eta_J$ is obtained from the others by the constraint that resource shares sum to one. This gives us one more demand equation to work with than we need for Theorem 1, which we can use to identify one more dimension of random variation. That is where the measurement error $e$ comes in. We assume the distribution of $e$ is independent of the true resource shares, which is reasonable since $e$ is measurement error and the variation in resource shares corresponds to true behavioural heterogeneity across households. With measurement error $e$, Theorem 1 uses a technique like that given above to identify the joint distribution of $\eta_1 e, ..., \eta_J e$ up to unknown locations (means), and Theorem 2 identifies those locations. We can then use $\sum_{j=1}^J \eta_j e = e$ to identify the distribution of $e$, and separate that out from the joint distribution of the resource shares $\eta_1, ..., \eta_{J-1}$.

### 3.6 Formal Identification

In this section we provide formal assumptions and our associated identification theorems. Proofs are in the Appendix.

**ASSUMPTION A1:** For every individual $j \in \{1, ..., J\}$ in the household there is a private, assignable good, denoted as good $j$. The household’s demand function for good $j$ is given by $Y_j = h_j(\eta_j e, x, p, z)$. The unknown functions $h_1, ..., h_J$ are differentiable and strictly increasing in their first element. Resource shares $\eta_1, ..., \eta_J$ are random variables having some unknown
joint distribution across households, with $\sum_{j=1}^{J} \eta_j = 1$. The measurement error $e$ is also a random variable that has an unknown distribution across households.

As discussed earlier, the assumptions that $Y_j = h_j(\eta_j, x, p, z)$ with $\sum_{j=1}^{J} \eta_j = 1$ follow immediately from the general framework of the Pareto efficient collective household model, with the goods indexed by $j$ being private and assignable (other goods may be shared and non assignable). The assumption that each $h_j$ is increasing in its first element just means that good $j$ is a normal good, i.e., a good for which demand goes up when total expenditures goes up.

Distribution factors $d$ are defined to be characteristics that affect $\eta_j$ but not $h_j$. Previous collective household models assumed that each $\eta_j$ is a deterministic function of $d$ and other observed variables. In contrast, we assume that $\eta_j$ varies randomly across households, or equivalently, that there exist unobserved distribution factors. This variation in each $\eta_j$ induces variation in observed private assignable goods demands $Y_j$.

For now we are assuming the only source of random variation across households are the resource shares $\eta_j$, but later we will add additional random variation that could be due to preference heterogeneity or measurement errors in $Y_j$.

Let $F_{Y_j}(Y_1, \ldots Y_j | p, x, d, z)$ denote the joint distribution of expenditures on private assignable goods $Y_1, \ldots Y_j$, conditioning on $p, x, d,$ and $z$.

**ASSUMPTION A2:** $F_{Y_j}(Y_1, \ldots Y_j | p, x, d, z)$ is identified from data for all $p, x, d,$ and $z$ in some sets $\Phi_p, \Phi_x, \Phi_d,$ and $\Phi_z$, respectively. The set $\Phi_x$ is an interval, and the set $\Phi_p$ is not empty.

A sufficient but stronger than necessary condition for Assumption A2 to hold is that we have iid observations of standard household survey data $(Y_1, \ldots Y_j, p, x, d, z)$ where the support of $p, x, d, z$ includes $\Phi_p \times \Phi_x \times \Phi_d \times \Phi_z$. But Assumption A2 could alternatively hold under many other standard sampling schemes, such as repeated cross section data where prices only vary by time, or situations where there is weak dependence in the observations, such as mixing. The set $\Phi_z$ could be empty, corresponding to not observing any demographic or other characteristics $z$ that might affect tastes. The set $\Phi_p$ could just contain a single element, in which case we will have Engel curve data with no price variation, as in a single cross section of households. Having $\Phi_x$ contain an interval implies that observed households have some continuous variation in total expenditures $x$, though the range of observed total expenditure values $x$, given by the interval $\Phi_x$, could be very small.

**ASSUMPTION A3:** Assume that $\eta_1, \ldots \eta_J$, conditional on any $p \in \Phi_p, d \in \Phi_d, z \in \Phi_z, x \in \Phi_x,$ and any $e$, has a continuous distribution that is independent of $x$ and $e$. Let
$F_{\eta}(\eta_1, ..., \eta_J \mid p, d, z)$ denote the unknown joint distribution of $\eta_1, ..., \eta_J$ conditional on $p, d, z$.

Lewbel and Pendakur (2008) and Bargain and Donni (2009) assume $e = 1$ and $\eta_1, ..., \eta_J$ independent of $x$ as in Assumption A3 to obtain identification in the case of deterministic rather than random resource shares. BCL and Lise and Seitz (2011) imposed $\eta_1, ..., \eta_J$ independent of $x$ on their empirical models (again having deterministic shares). DLP and Menon, Pendakur, and Perali (2012) provide both theoretical and empirical evidence supporting this independence assumption, and Cherchye, De Rock, Lewbel and Vermeulen (2015) find that identified bounds on resource shares also support this assumption. Note that this independence only needs to hold after conditioning on observables $z$ that can include demographic characteristics and observable distribution factors. One way to interpret Assumption A3 is to assume there exist unobservable distribution factors, including at least one that is continuously distributed, and that the distribution of unobserved distribution factors across households does not depend on $x$, after conditioning on $p$ and $z$.

ASSUMPTION A4: Assume that $e > 0$ is independent of $\eta_1, ..., \eta_J, p, d, x$, and $z$, that $E(e) = 1$, and that $\eta_1, ..., \eta_{J-1}, e$ has a nonvanishing characteristic function.

The assumption that $E(e) = 1$ means that $E(x^*) = E(x)$, corresponding to the standard measurement error assumption that the observed variable is on average correctly measured, so total expenditures are not systematically over or undercounted.

THEOREM 1: Let Assumptions A1, A2, A3 and A4 hold. Then, for some unknown functions $c_1(p, z), ..., c_J(p, z)$ the joint distribution of $\eta_1 ec_1(p, z), ..., \eta_J ec_J(p, z)$ conditional on $p, d, z$ is identified for all $p \in \Phi_p, d \in \Phi_d$ and $z \in \Phi_z$.

A well known nonidentification result in the collective household literature is that without restrictions on preferences, the levels of (deterministic) resource shares cannot be identified (see Chiappori and Ekelund 2009 for a current general version of this result). Instead, only changes in resource shares with respect to observed distribution factors can be identified. This is equivalent to saying that, if $e$ and each $\eta_J$ were deterministic functions of observables, then only $\eta_1 eC_1(p, z), ..., \eta_J eC_J(p, z)$ could be identified for some unknown functions $C_j(p, z)$. Theorem 1 provides a substantial generalization of this result to random resource shares. For example, consider the special case where $e$ is constant. Theorem 1 then says that when $\eta_j$ are random, the entire joint distribution of resource shares (and hence the effect of all unobserved distribution factors) is identified up to the same unknown deterministic functions $C_j(p, z) = ec_j(p, z)$. The distribution identification result in Chiappori and Kim
(2017) is also a special case of Theorem 1, and in particular follows from Theorem 1 when \( e \) is constant and \( J = 2 \). Note that Chiappori and Kim also require considerable price variation, while our results allow \( \Phi_p \) to have few or even just one element.

**ASSUMPTION A5:** Assume \( \Phi_d \) contains at least \( J \) elements, which without loss of generality will be denoted \( d_1, \ldots, d_J \). Assume \( E(\eta_j \mid p, d, z) \neq 0 \). Let \( T(p, z) \) be the \( J \) by \( J \) matrix defined by having \( E(\eta_j \mid p, d_k, z) / E(\eta_j \mid p, d_1, z) \) in the row \( k \) and column \( j \) position. Assume there exist sets \( \Phi_p^* \) and \( \Phi_z^* \), which are subsets of \( \Phi_p \) and \( \Phi_z \), such that for all \( p \in \Phi_p^* \) and \( z \in \Phi_z^* \), \( T(p, z) \) is nonsingular.

The key feature of Assumption A5 is the requirement that our set of distribution factors must take on at least \( J \) values, recalling that \( J \) is the number of household members. The nonsingularity of \( T(p, z) \) required by Assumption A4 will generally hold, failing only when there is some equality coincidence among the expected resource share functions \( E(\eta_j \mid p, d_1, z) \). For example, in households with two members, it is straightforward to check that nonsingularity will hold as long the distribution factor affects the mean of \( \eta_1 \) in any way, that is, as long as \( E(\eta_j \mid p, d_1, z) \neq E(\eta_j \mid p, d_2, z) \). Also, regarding \( E(\eta_j \mid p, d, z) \neq 0 \), we actually only require that this hold for one tuple of given known, observed values of \( p \), \( d \), and \( z \).

**THEOREM 2:** Let Assumptions A1, A2, A3, A4, and A5 hold. Then \( F_{\eta}(\eta_1, \ldots, \eta_J \mid p, d, z) \) is identified for all \( d \in \Phi_d, p \in \Phi_p^* \) and \( z \in \Phi_z^* \). The distribution of \( e \) is also identified.

Theorem 2 works by exploiting the fact that resource shares must sum to one within a household. For example, if \( \Phi_d \) has \( J \) elements, this places \( J \) equality constraints on the set of functions \( E(\eta_j \mid p, d, z) \), one for each of the \( J \) values that the distribution factors can take on. By Theorem 2, the \( E(\eta_j \mid p, d, z) \) are identified up to \( J \) unknown functions, and the \( J \) equality constraints allow us to recover these \( J \) unknown functions, \( c_1(p, z), \ldots, c_J(p, z) \). Given these \( c_j(p, z) \) functions, by Theorem 2 the entire joint distribution of the resource shares is identified.

Note that Theorems 1 and 2 do not require any price variation, and so can be applied to Engel curve type data where all observations are drawn from a single price regime.

An immediate extension of Theorem 2 is that we could have identified the levels of deterministic resource shares, e.g., in traditional nonstochastic collective household models, given resource shares independent of total expenditures and observing some distribution factors. Theorem 2 thereby overcomes the classic resource shares nonidentification problem. So, e.g., the SAP and SAT preference restrictions employed by DLP could be replaced with Assumption A4.
4 Semiparametric Estimation

Although we have established nonparametric identification of the model, completely nonparametric estimation will often be impractical with typical sized data sets, given the high dimensionality of the model. We therefore focus on estimation assuming a parametric specifications for the private assignable demand equations and for the deterministic component of resource shares, but nonparametric in the random variation in resource shares. From the long literature on empirical demand estimation, a great deal is known about functional forms for demand functions that are empirically successful. However, almost nothing is known about the distribution of random variation in resource shares. We therefore employ commonly used functional forms for utility, while leaving the distribution of resource shares (and measurement error $e$) unspecified and nonparametric.

Theorem 2 nonparametrically identifies the distribution of random resource shares (and an associated estimator is provided in the Appendix), but Theorem 2 does not identify the separate resource shares of each household. In this section we make stronger functional form assumptions that will permit estimation of resource shares for each household. The empirical distribution of these individual resource shares can then provide an estimate of the distribution of resource shares. However, unlike Theorem 2, we do not have a general theorem describing when resource shares for each household can be identified and estimated. Rather, we here just discuss estimation under the specific model we bring to the data.

Our theoretical model assumes a separate utility function and a separate assignable goods for each household member. In our empirical application we take men’s, women’s, and children’s clothes and shoes to be our private assignable goods $y_j$. Due to data limitations we are assuming a single utility function for all the children in the household. This utility function includes a children’s assignable good, which is children’s clothing and shoes (DLP make the same assumption for the same reason). If we had observed separate clothing and or shoes expenditures for each child, we could have allowed each to have their own utility function. We therefore have $j$ taking three different values, corresponding to the three types of household members: father, mother, and children.

Let $\theta_j$ be a vector of clothing demand parameters. Other variables are defined as before. Based on our theorems, assume $(\eta_1, \eta_2, \eta_3, e) \perp x | z, p, d$ and let $E(e | \eta_1, \eta_2, \eta_3, z, p, d) = 1$. The general form of a demand function is

$$y_j = h_j(x\eta_j e, p, z | \theta_j)$$

where the distribution of each resource share $\eta_j$ can depend on $z, p, d$. In the appendix we provide a general technique based on inverse demands for estimating $\theta_j$ and the values of
η₁, η₂, η₃, e for each individual in the sample. Below we give a simpler estimator based on our specifically chosen functional form for h_j.

Our empirical application uses cross-sectional data from a single time period, so we assume that p is fixed, i.e., we work with Engel curve data. We can then drop p from the above demand equation, and let each y_j just equal observed expenditures on member j’s clothing. The estimator described below will allow us to directly estimate the mean value of each η_j conditional on covariates, and then later back out estimated values of these resource shares for each j in each household in our sample.

We assume that the Marshallian demand function h_j for the private assignable good has the commonly used piglog (price independent generalised logarithmic) functional form. The name piglog is due to Muellbauer (1976). This is one of the most widely used specifications for Engel curves, e.g., the Jorgenson, Lau, and Stoker (1982) Translog demand system and the Deaton and Muellbauer (1980) Almost Ideal Demand System have Engel curves of the piglog form, and piglog Engel curves were also used in empirical collective household models estimates by DLP and by Calvi (2017). Note that only the assignable goods are assumed to be piglog, so more general specifications could apply to other goods, such as the QUAIDS model of Banks, Blundell, and Lewbel (1997).

The piglog Engel curve model is

\[ h_j(x, z | \theta_j) = c_j(z) x + b_j(z) x \ln(x) \]  \tag{6} 

where c_j(z) and b_j(z) are implicitly functions of parameters \( \theta_j \). The corresponding household demand function for each private assignable good j is

\[ y_j = c_j(z) x \eta_j e + b_j(z) x \eta_j e \ln(x \eta_j e) \]

Let \( w_j = y_j / x \) be the budget share of person j’s assignable good as a fraction of total household expenditures x. We then have that the household’s budget-share function for the assignable good j is

\[ w_j = [c_j(z) + b_j(z) \ln(x \eta_j e)] \eta_j e \]

As noted above, due to data limitation we treat all children in the household as a single

\[ \text{Piglog demands are given by the class of indirect utility functions } V(\ln P, \ln x, z) = (\ln x - A(P, z)) \exp(B(P, z)) \text{ for some functions } A(P, z) \text{ and } B(P, z), \text{ where } \exp(A(P, z)) \text{ is homogeneous of degree one in } P \text{ and } B(P, z) \text{ is homogeneous of degree zero in } P. \text{ Applying Roys identity to this indirect utility gives the piglog demand functions, in which the budget share of good } j \text{ equals } A_j(P, z) - A(P, z) B_j(P, z) - B_j(P, z) \ln x, \text{ where } A_j(P, z) \text{ is the partial derivative of } A(P, z) \text{ with respect the price of good } j, \text{ and similarly for } B_j(P, z). \text{ Holding prices fixed at some constant level } P_0, \text{ define } \text{b}_j(z) = -B_j(P_0, z) \text{ and } c_j(z) = A_j(P_0, z) - A(P_0, z) B_j(P_0, z) \text{ to obtain equation (6).} \]
household member type. Let $N_j$ denote the number of household members of each type $j$. In our data, $N_j$ equals one for each adult type ($j$ being father or mother), and equals 1, 2, 3, or 4 for the number of children. Adjusting the demand equation to allow for multiple members of any type $j$ gives

$$w_j = [c_j(z) + b_j(z) \ln (x\eta_j e/N_j)] \eta_j e$$

Define $\mu_j$ to be the conditional mean function for $\eta_j$, given by

$$\mu_j (d, z) = E(\eta_j e \mid x, d, z) = E(\eta_j \mid d, z).$$

Define $U_j$ to be a structural error term, which depends on the deviation of $\eta_j$ from $\mu_j$ as

$$U_j = \frac{\eta_j e}{\mu_j (d, z)}.$$

If resource shares $\eta_j$ are deterministic functions of observables, then $\mu_j (d, z) = \eta_j$, so in that case $\mu_j$ are the resource shares. Otherwise, if resource shares $\eta_j$ are random (functions of unobserved distribution factors $v$) or if measurement errors $e$ are present then $U_j$ is the random variation in $\eta_j e$ across individuals. Substituting these expressions into the above household budget-share functions gives

$$w_j = (c_j(z) + b_j(z) (ln (x\mu_j (d, z) /N_j)) + \ln U_j)) \mu_j (d, z) U_j$$

Gathering $U_j$ terms this becomes

$$w_j = [c_j(z) + b_j(z) \ln (x\mu_j (d, z) /N_j)] \mu_j (d, z) U_j + b_j(z) \mu_j (d, z) U_j \ln U_j.$$ 

(7)

Since $E(e \mid \eta_1, \eta_2, \eta_3, z, p, d) = 1$ and $E(\eta_j \mid d, z) = \mu_j (d, z)$, we have that $E(U_j \mid x, d, z) = 1$. Let $s_j = E(U_j \ln U_j \mid x, d, z)$, which for simplicity is assumed to be constant (letting $s_j$ depend on covariates is feasible, but complicates the numerical evaluations below). This is roughly analogous to a simplifying homoscedasticity assumption. Define an additive error term $\varepsilon_j$ by $\varepsilon_j = w_j - E(w_j \mid x, d, z)$, and therefore $E(\varepsilon_j \mid x, d, z) = 0$.

Define $C_j(z)$ by

$$C_j(z) = c_j(z) + b_j(z) s_j$$

Evaluating $E(w_j \mid x, d, z)$ from equation (7) gives

$$w_j = [C_j (z) + b_j(z) \ln (x\mu_j (d, z) /N_j)] \mu_j (d, z) + \varepsilon_j$$

(8)
We can think of the $U_j$’s as structural errors, while the $\varepsilon_j$’s are ordinary regression errors. We will assume for estimation that $C_j(z)$, $b_j(z)$, and $\mu_j(d,z)$ are all linear in their arguments, so

$$
\mu_j(d,z) = \alpha_{j0} + \alpha'_{jz}z + \alpha'_{jd}d, \quad b_j(z) = \beta_{j0} + \beta'_{jz}z \quad \text{and} \quad C_j(z) = \gamma_{j0} + \gamma'_{jz}z
$$

for constant scalars $\alpha_{j0}$, $\beta_{j0}$, and $\gamma_{j0}$, and constant coefficient vectors, $\alpha_{jz}$, $\alpha'_{jd}$, $\beta_{jz}$, and $\gamma_{jz}$. Substituting these expressions into equation (8) gives the model that we estimate, which for each member $j$ of each household $h$, is

$$
\begin{align*}
w_{jh} &= \left[ \gamma_{j0} + \gamma'_{jz}z_h + (\beta_{j0} + \beta'_{jz}z_h) \left[ \ln \left( \frac{x_h}{N_{jh}} \right) + \ln \left( \alpha_{j0} + \alpha'_{jz}z_h + \alpha'_{jd}d_h \right) \right] \right] \left( \alpha_{j0} + \alpha'_{jz}z_h + \alpha'_{jd}d_h \right) + \varepsilon_{jh}
\end{align*}
$$

Equation (9) for household members $j = 1, \ldots, J$ now form a system of $J$ equations that we can estimate using a least squares estimator such as nonlinear seemingly unrelated regression. We will also want to employ clustered (across equations), heteroskedastic robust standard errors, to account for the fact that, by construction, the errors $\varepsilon_j$ are heteroskedastic and are correlated across the members of each household.

An interesting feature of this model and our associated estimator is that the estimating equations (9) for each $j$ are the same regardless of whether resource shares are random or not. If resource shares are not random and measurement errors $e = 1$ then $s_j = 0$ would make $c_j(z)$ equal $C_j(z)$ and make $\eta_{jh}$ equal $\mu_j(d_h, z_h)$, where $\eta_{jh}$ is the resource share of member $j$ in household $h$. But if resource shares are random or measurement errors $e$ are present, then (given the parameter estimates) we can use the fact that the errors $\varepsilon_{jh}$ are known functions of the random components $U_{jh}$ to recover random resource shares for each household, as follows.

Define $\psi_{0jh}$ and $\psi_{1jh}$ by

$$
\begin{align*}
\psi_{0jh} &= [C_j(z_h) + b_j(z_h) \ln (x \mu_j(d_h, z_h) / N_{jh})] \mu_j(d_h, z_h) \\
\psi_{1jh} &= b_j(z_h) \mu_j(d_h, z_h).
\end{align*}
$$

Note that $\psi_{0jh}$ and $\psi_{1jh}$, are known functions of parameters to be estimated. Then we can rewrite equations (8) and (7) more simply as

$$
\begin{align*}
w_{jh} &= \psi_{0jh} + \varepsilon_{jh} \quad \text{and} \quad w_{jh} = (\psi_{0jh} - \psi_{1jh}s_j)U_{jh} + \psi_{1jh}U_{jh} \ln U_{jh}
\end{align*}
$$

For each $j$ in each household $h$, we observe $w_{jh}$ and have estimates of $\psi_{0jh}$ and $\psi_{1jh}$. For each $j$, we determine $s_j$ and $U_{j1}, \ldots, U_{jh}$ by numerically minimizing the difference between
$s_j$ and the sample average of $U_{jh} \ln U_{jh}$ plus a penalty function, subject to the constraint that equation (10) hold for each $U_{jh}$.\textsuperscript{4,5} The penalty function penalizes departures of the sample average of $U_{jh}$ from one. The purpose of this penalty, which is based on the model restriction that $E(U_j \mid x, d, z) = 1$, is to choose the proper solutions in the event that the minimization has multiple solutions.\textsuperscript{6} We do not have a formal proof that this procedure will always converge. However, in our empirical application even with widely varying magnitudes of the penalty factor we did not obtain multiple solutions, the procedure always converged, and we obtained almost identical answers regardless of the chosen penalty factor.

Once we have obtained each person’s $U_{jh}$ for each type $j$ in each household $h$ we calculate

$$
e_h = \sum_{j=1}^{J} U_{jh} \mu_j (d_h, z_h) \quad \text{and} \quad \eta_{jh} = \frac{U_{jh} \mu_j (d_h, z_h)}{e_h}$$

where $e_h$ is the multiplicative measurement error in $x_h$ for household $h$ and $\eta_{jh}$ is the random resource share of person $j$ in household $h$. These expressions follow from the constructions $U_{jh} = \eta_{jh} e_h / \mu_j (d_h, z_h)$ and $\sum_{j=1}^{J} \eta_{jh} = 1$.

5 Empirical Results

5.1 Malawian Expenditure Data

Our data are from the fourth Malawi Integrated Household Survey (IHS4).\textsuperscript{7, 8} The IHS4 provides detailed Malawian household expenditure and demographic data collected in 2016/2017.

\textsuperscript{4}We implement this solution numerically using the constrained minimization routine in MatlabR2018b.

\textsuperscript{5}For households that have $\mu_{jh}$ very close to zero, the resulting value of $U_{jh}$ becomes very large, resulting in very large sample values of $U_{jh} \ln U_{jh}$ which could destabilise the numerical search. To avoid this problem, we trimmed the data by dropping observations where $\mu_{jh} < 0.01$ for any household member (this entailed dropping 5 of our 2887 households).

\textsuperscript{6}For each $j$, We have $H+1$ equations in $H+1$ unknowns, however, equation (10) can have up to two solutions for each $U_{jh}$. The correct solutions must asymptotically satisfy $s_j = E(U_j \ln U_j)$ and $E(U_j \mid x, d, z) = 1$.

\textsuperscript{7}IHS4 was conducted in 2016-2017 by the National Statistics Office of the Government of Malawi under the World Bank Living Standards Measurement Survey - Integrated Surveys on Agriculture initiative. The survey sampled 12,480 households and is designed to be statistically representative at the national, district, urban and rural levels for Malawi. The IHS4 is a cross-sectional survey though sister surveys on a panel of households are also available. The IHS4 microdata is publicly available from the World Bank website: http://micrdata.worldbank.org/index.php/catalog/2939.

\textsuperscript{8}Malawi is a very poor former British protectorate in southern Africa. With a population density of approximately 210 persons per sq. km, it is one of the most densely populated countries in Africa, even though roughly 83% of the population lives in rural areas. Unlike many poor African countries, Malawi’s recent political history is remarkable for the absence of military coups and for occasional multi-party elections, most recently in 2009. Despite its relative political stability, Malawi has numerous socio-economic tensions including extreme poverty, a high incidence of HIV/AIDS, high infant mortality and one of the lowest life expectancies in the world (64 years).
by the National Statistics Office of Malawi in partnership with the World Bank\textsuperscript{9}. These data are similar to that used DLP except that they are from a more recent vintage of the same household survey and include more variables.

While household level poverty rates are quite high in Malawi, we find that accounting for intra-household allocations results in even higher rates of poverty. In particular, our results show many people (particularly children) are individually below a poverty line, despite living in households where everyone in the household could be above the line if intra-household resources were distributed more equitably.

In the household module, households are asked questions from a number of submodules relating to health, education, employment, fertility and consumption. Households are asked to recall their food consumption (one week recall) and their non-food expenditure broken into four recall categories (one week, one month, three months and one year). In our work, we convert consumption to the annual level (e.g., multiplying one month recall consumption by 12). Consumption amounts also include the value of home produced goods and services imputed at the value of those services consumed in the market.

The consumption data include (in the three month recall questionnaire) household expenditures on clothing and shoes separately for the household head, spouse(s), and children aged 14 and under. These are the assignable goods whose demand functions we model. We use a single private assignable good for each person equal to the sum of clothing and footwear expenditures for that person. As noted above, we treat “children” as an aggregate person.

Our new method for identifying the level of resource shares depends on observing distribution factors, defined as variables that affect resource shares but do not affect individual preferences. Browning, Chiappori and Weiss (2010) suggest a number of plausible distribution factors that might be found in microdata. One particular advantage of the IHS4 data is that it contains two of the distribution factors that were suggested by Browning, Chiappori and Weiss (2010).\textsuperscript{10} We construct distribution factors using community-level information from the community module.

\textsuperscript{9}These are relatively high quality data. Enumerators were recruited using national newspaper advertisements and successful candidates were given training by the IHS4 Management Team and World Bank specialists. Enumerators were sent to individual households to collect the data for IHS4. Enumerators were assigned to survey teams and monitored by Field Supervisors in order to ensure that the random samples were followed and also to ensure high data quality. Completed surveys were reviewed by a headquarters team and call back interviews were assigned for questionnaires that failed the review process. Survey review was timely because Enumerators used Samsung Galaxy Tab S2 tablets to administer the surveys and Field Supervisors used laptop computers and wireless routers to sync completed data to Headquarters. Households were selected for the survey by the headquarters’ team to ensure that the final sample of households was not biased by survey teams in field choices. Enumerators were responsible for collecting data for three household-level modules: household, agriculture and fisheries; and one community-level module.

\textsuperscript{10}Previous versions of this paper used IHS2, which did not contain these recommended distribution factors.
Our first distribution factor is the natural log of the ratio of the female and male ganyu wage for labourers, measured at the community level as reported by village elders. Ganyu labour is a short-term, mainly rural, labour relationship that is a key source of income for poor households in Malawi. The relative ganyu wage is therefore a measure of the potential relative wage for each household member. The relative wage is a continuous variable in the data as it is neither top-coded nor reported in ranges. Notably, this wage ratio is available for all households, regardless of whether or not either or both spouses are working for wages, so it provides a measure of the ratio of opportunity wages for men and women. Consequently, this measure neither pollutes the sample with selection bias (e.g., from including only two-worker households) nor household-level wage heterogeneity.

Our second distribution factor is a binary variable that equals one if households in the community trace their descent through their father, and equals zero if descent is traced either through the mother or both parents. Tracing descent through the father implies that, by custom, fathers have control over any post-marriage assets. Roughly 20 per cent of households in our sample reside in villages that trace descent through the father.\textsuperscript{11}

We use the relative wage and the binary descent variable as the distribution factors $d$ in our structural models. Since both distribution factors are measured at the community level, we cluster our reported asymptotic standard errors at the community level.

We provide a number of tests and other evidence that the relative wage and binary descent variables satisfy the properties required of distribution factors. First, as we demonstrate below, they are statistically significant in resource shares functions. Second, to support the requirement that they not affect preferences, we first note that both of our distribution factors are community-level and so do not depend on the exact characteristics of the households sampled (because the sampled households are only a small part of the total village populations). Third, we can and do directly test whether or not these candidate distribution factors affect preferences given the similar across people (SAP) preference restriction of DLP, because SAP provides overidentifying restrictions. In particular, we find that they can be excluded from preferences: the sample value of the chi-squared test statistic for the hypothesis that our two distribution factors can be excluded from the preference functions $\gamma_j$ and $\beta$ is equal to 6.4 with 8 degrees of freedom with a p-value of 0.61.

Our analysis includes 16 demographic variables, denoted $z$, which may affect both preferences and resource shares of each household member. These variables are: region of residence (non-urban North and non-urban Central with non-urban South as the left-out category);

\textsuperscript{11}This question is related to matrilineal and patrilineal marriage, e.g. Walther (2018), but it is distinct. Marriage type is also a question in the IHS4 survey but it differentiates between matrilineal/patrilineal and matrilocal/patrilocal and neolocal which suggests that marriage type may confound the location of post-marriage residence and the location of marriage.
the average age of children less 5; the minimum age of children less 5; the proportion of children who are girls; the age of the man less 28 and the age of the woman less 22 (the average ages of men and women in the sample); the education levels of the household head and spouse (ranging from $-2$ to 4, where 0 is the modal education level); the log of the distance of the village to a road and to a daily market; and dummy variables indicating that the household is christian or muslim (with animist/other as the left-out category). These variables are constructed so that setting $z = 0$ corresponds to a typical or average household in the data.

Additionally, we include the number of children, $s$, among our demographic variables, in the form of dummy variable indicators for 2-, 3- and 4-child households. These are the same demographic variables as used in DLP except that we do not include a dry-season dummy (this variable was not included in the IHS4 data). Thus, $z$ has 16 elements: 13 demographic elements and 3 elements equal to values of $s$.

Table 1: Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std dev</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>clothing shares</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>man</td>
<td>0.0048</td>
<td>0.0119</td>
<td>0.0000</td>
<td>0.1596</td>
</tr>
<tr>
<td>woman</td>
<td>0.0085</td>
<td>0.0144</td>
<td>0.0000</td>
<td>0.1734</td>
</tr>
<tr>
<td>children</td>
<td>0.0124</td>
<td>0.0183</td>
<td>0.0000</td>
<td>0.1961</td>
</tr>
<tr>
<td>(left-out)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s=1</td>
<td>0.3055</td>
<td>0.4607</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>s=2</td>
<td>0.3027</td>
<td>0.4595</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>s=3</td>
<td>0.2508</td>
<td>0.4335</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>s=4</td>
<td>0.1410</td>
<td>0.3481</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d1: ln relative wage</td>
<td>-0.1499</td>
<td>0.2932</td>
<td>-1.0986</td>
<td>1.0986</td>
</tr>
<tr>
<td>factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d2: patrilineal</td>
<td>0.2245</td>
<td>0.4173</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>


We use almost the same sample restrictions as DLP. As in that paper, our sample here consists of married couples with one to four children all under 15 years of age that satisfy the following sample restrictions: (1) polygamous marriages are excluded; (2) observations with any missing data for regressors are excluded; (3) households with children aged 15 or over are excluded; (4) households with any member over 65 are excluded; (5) households with nonrelatives members are excluded; and (6) urban households are excluded.

Our final sample contains 2887 households. Table 1 gives some summary statistics. Note that the log relative Ganyu wage is less than zero, showing that female wages are lower than male wages in rural Malawi.
5.2 Estimation

We estimate equation (9), by nonlinear seemingly unrelated least squares.\textsuperscript{12} To account for variation in the number of children across households, in equation (9) we replace $\ln x_h$ in the childrens’ assignable good equation with $\ln (x_h/s)$. The two distribution factors $d$ are included in all models.

In the main text tables, we report parameters $\alpha_{j0}$, $\alpha_{js}$, $\beta_{j0}$ and $\beta_{js}$ where $s = 2, 3, 4$. So, e.g., $\alpha_{j0}$ and $\beta_{j0}$ give the resource share and budget response, respectively, for a household with $z = 0$. This household has 1 child (the left-out household size) and all demographic variables equal to 0. The reported parameters $\alpha_{js}$ and $\beta_{js}$ give the difference in the impact on the resource share and budget response, respectively, of having 2, 3 or 4 children, versus one child. Values for the children’s resource share parameters are calculated (as one minus the sum of adults’ shares) with standard errors obtained via the delta-method. For resource shares, we also report parameter $\alpha_{jd}$, $d = 1, 2$, which give the marginal response of resource shares to the two distribution factors.

We provide estimates where identification hinges on the existence of observed distribution factors as in Theorem 1, denoted “DF” in the tables. We also provide estimates imposing the similar across people restrictions of DLP, denoted “SAP” in the tables. SAP imposes the constraint that $\beta_{js} = \beta_s$ for $s = 0, 2, 3, 4$.

We estimate the models via nonlinear seemingly unrelated regression (\texttt{nlsur}) using Stata 15. Even though the model is only slightly nonlinear, we found empirically that there were multiple local minima. To find the global minimum, we generated a set of 250 random starting values for $\eta_j$ and $\beta$ for the \texttt{nlsur} estimator and selected the estimates with the lowest observed minimized sum-of-squared residuals.\textsuperscript{13}

The next subsection reports our estimates of the levels of resource shares. This is followed by estimates of the random distribution of resource shares around these levels, and then the resulting estimates of separate poverty rates for men, women, and children.

\textsuperscript{12}This estimator is asymptotically equivalent to the Generalized Method of Moments (GMM), and equivalent to efficient GMM if the errors are homoscedastic. We report clustered standard errors, with clusters at the village level, thereby providing asymptotically valid inference even if correlations exist in the errors across households within the same village.

\textsuperscript{13}We used Stata 15 on Windows 10 OS to implement the search for the best SAP estimates. We used MATLAB on Windows 10 OS to implement the inversion (described below) that recovers $U_{jh}$ from $e_{jh}$. We used Stata 15 on MAC OS to estimate DF estimates from SAP starting values, and to construct all tables below. We noticed empirically that DF estimates using starting values distant from the best SAP estimates were numerically unstable and had many local minima. However, DF estimates using best SAP estimates as starting values were unique.
5.3 Estimated Coefficients and Levels of Resource Shares

Table 2 gives estimates of selected elements of resource share parameters vector \((\alpha)\) and the budget response parameter vector \((\beta)\) for our model, and for the more restrictive SAP model of DLP (Table A1 in the Appendix provides the full set of estimates of all parameters). At the bottom of the SAP block, we provide a test of the restrictions \(\beta_{js} = \beta_s\), which is a test of SAP subject to the validity of the distribution factors.

<table>
<thead>
<tr>
<th>function</th>
<th>variable</th>
<th>SAP est</th>
<th>std err</th>
<th>DF est</th>
<th>std err</th>
</tr>
</thead>
<tbody>
<tr>
<td>men’s</td>
<td>constant</td>
<td>0.3145</td>
<td>0.0428</td>
<td>0.3669</td>
<td>0.0770</td>
</tr>
<tr>
<td>resource</td>
<td>s=2</td>
<td>-0.0255</td>
<td>0.0345</td>
<td>-0.0488</td>
<td>0.0336</td>
</tr>
<tr>
<td>share</td>
<td>s=3</td>
<td>-0.1001</td>
<td>0.0410</td>
<td>-0.1285</td>
<td>0.0456</td>
</tr>
<tr>
<td>((\alpha_m))</td>
<td>s=4</td>
<td>-0.0804</td>
<td>0.0527</td>
<td>-0.1274</td>
<td>0.0531</td>
</tr>
<tr>
<td></td>
<td>d1</td>
<td>0.0541</td>
<td>0.0170</td>
<td>0.0529</td>
<td>0.0178</td>
</tr>
<tr>
<td></td>
<td>d2</td>
<td>0.0249</td>
<td>0.0139</td>
<td>0.0252</td>
<td>0.0143</td>
</tr>
<tr>
<td>woman’s</td>
<td>constant</td>
<td>0.3834</td>
<td>0.0551</td>
<td>0.3840</td>
<td>0.0816</td>
</tr>
<tr>
<td>resource</td>
<td>s=2</td>
<td>-0.0313</td>
<td>0.0343</td>
<td>-0.0218</td>
<td>0.0346</td>
</tr>
<tr>
<td>share</td>
<td>s=3</td>
<td>-0.0356</td>
<td>0.0479</td>
<td>-0.0358</td>
<td>0.0491</td>
</tr>
<tr>
<td>((a_f))</td>
<td>s=4</td>
<td>-0.0461</td>
<td>0.0617</td>
<td>-0.0437</td>
<td>0.0652</td>
</tr>
<tr>
<td></td>
<td>d1</td>
<td>-0.0055</td>
<td>0.0130</td>
<td>-0.0092</td>
<td>0.0138</td>
</tr>
<tr>
<td></td>
<td>d2</td>
<td>-0.0211</td>
<td>0.0111</td>
<td>-0.0238</td>
<td>0.0115</td>
</tr>
<tr>
<td>children’</td>
<td>constant</td>
<td>0.3020</td>
<td>0.0573</td>
<td>0.2490</td>
<td>0.0687</td>
</tr>
<tr>
<td>resource</td>
<td>s=2</td>
<td>0.0569</td>
<td>0.0344</td>
<td>0.0706</td>
<td>0.0322</td>
</tr>
<tr>
<td>share</td>
<td>s=3</td>
<td>0.1358</td>
<td>0.0462</td>
<td>0.1642</td>
<td>0.0498</td>
</tr>
<tr>
<td>((a_c))</td>
<td>s=4</td>
<td>0.1265</td>
<td>0.0584</td>
<td>0.1710</td>
<td>0.0583</td>
</tr>
<tr>
<td></td>
<td>d1</td>
<td>-0.0487</td>
<td>0.0172</td>
<td>-0.0437</td>
<td>0.0182</td>
</tr>
<tr>
<td></td>
<td>d2</td>
<td>-0.0038</td>
<td>0.0135</td>
<td>-0.0013</td>
<td>0.0128</td>
</tr>
<tr>
<td>slope</td>
<td>constant</td>
<td>0.0159</td>
<td>0.0039</td>
<td>0.0089</td>
<td>0.0048</td>
</tr>
<tr>
<td>((\beta_m))</td>
<td>s=2</td>
<td>-0.0032</td>
<td>0.0033</td>
<td>-0.0022</td>
<td>0.0039</td>
</tr>
<tr>
<td></td>
<td>s=3</td>
<td>-0.0027</td>
<td>0.0041</td>
<td>-0.0006</td>
<td>0.0046</td>
</tr>
<tr>
<td></td>
<td>s=4</td>
<td>-0.0035</td>
<td>0.0048</td>
<td>0.0006</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>constant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>((\beta_f - \beta_m))</td>
<td>constant</td>
<td>0.0093</td>
<td>0.0076</td>
<td></td>
</tr>
<tr>
<td></td>
<td>s=2</td>
<td>-0.0019</td>
<td>0.0045</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>s=3</td>
<td>-0.0031</td>
<td>0.0060</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>s=4</td>
<td>-0.0097</td>
<td>0.0076</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>((\beta_c - \beta_m))</td>
<td>constant</td>
<td>0.0109</td>
<td>0.0094</td>
<td></td>
</tr>
<tr>
<td></td>
<td>s=2</td>
<td>-0.0016</td>
<td>0.0077</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>s=3</td>
<td>-0.0047</td>
<td>0.0104</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>s=4</td>
<td>-0.0044</td>
<td>0.0124</td>
<td></td>
<td></td>
</tr>
<tr>
<td>test dist facts</td>
<td>chi2, pval</td>
<td>16.5</td>
<td>0.0024</td>
<td>15.3</td>
<td>0.0041</td>
</tr>
<tr>
<td>test SAP</td>
<td>chi2, pval</td>
<td>29.7</td>
<td>0.6790</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the upper panels of the left-hand block, we show selected estimates of resource share parameters given SAP. Consider first the estimated values for the number of children. The upper left coefficient is $\alpha_{\text{man},0}$, giving the estimated level of the resource share for a man in a household with 1 child and all other demographic variables and distribution factors equal to 0. The estimate indicates that the man in such a household has a shadow budget equal to 0.3145 times the overall household budget $x_h$, so his resource share is about 31.5%. The estimate of 0.3834 for $\alpha_{\text{woman},0}$ indicates that the woman in such a household has a resource share of 38.3%, leaving 30.2% of the household budget for the child.

As in DLP, we find that households with more children have lower resource shares for parents. However, unlike DLP, we find that the marginal impact of increasing the number of children is larger for men than for women. For example, in households with 3 children, men’s resource shares are 10 percentage points lower, and women’s resource shares are 4 percentage points lower, than in households with 1 child. This implies that children’s resource shares are 14 percentage points higher in those households.

Turning to the distribution factors, we find that they are important drivers of resource shares. The ratio of women’s to men’s ganyu wage has a statistically significantly positive effect on male resource shares and a statistically significantly negative effect on children’s resource shares (with no significant effect on women’s resource shares). That is, relatively higher available wages for women in the village are, perhaps surprisingly, associated with a shift in resource shares from children to men. The second distribution factor is an indicator that the village of residence is primarily patrilineal in terms of post-marital asset division. Here, we see a statistically significant shift of resources from women to men (with no significant effect on children’s resource shares) associated with living in a patrilineal village.

The sample value of the test statistic for the hypothesis that these two distribution factors can be excluded from the resource shares is 16.5 and has a p-value of 0.0024. This suggests that our two distribution factors are highly relevant.

Next consider the slope parameters $\beta_{js}$. We find that the estimated values of the parameter $\beta_j$ is statistically significantly positive, and that the estimated values of $\beta_{j0} + \beta_{js}$ are all significantly positive for $j = 2, 3, 4$. These estimates indicate that clothing is a luxury for all household sizes. These estimates also indicate that a necessary condition for identification of resource shares (nonzero slope coefficients) is satisfied.

The right-hand panel of Table 2 shows selected parameter estimates relaxing the SAP restriction of SAP that $\beta$ is the same for all $j$. Here, identification requires that $\beta_j$ is nonzero for all $j$, and that the distribution factors are relevant. The estimated resource share parameters are quite similar to the estimates that impose SAP, but the estimated standard errors are much larger when the restrictions of SAP are relaxed. For example, the estimated
women’s resource share for $z = 0$ (one-child household) is 38 per cent given either model, but the estimated standard error is 5.5 percentage points given SAP and 8.2 percentage points when SAP is relaxed and the distribution factors do all the work of identification.

Imposing vs relaxing SAP has little effect on the values and significance of the distribution factors’ coefficients. The sample value of the test statistic for the hypothesis that these two distribution factors can be excluded from the resource shares is now 15.3 and has a p-value of 0.0041, again showing that the two distribution factors are jointly relevant.

The estimated values of $\beta_m$ are broadly similar to those given SAP. The estimated values of $\beta_f - \beta_m$ and of $\beta_c - \beta_m$ (the differences between $\beta$ across people) are individually insignificantly different from zero, supporting the SAP assumption. The bottom row of the table provides the sample value of the test statistic for the hypothesis that $\beta_f - \beta_m$ and $\beta_c - \beta_m$ are both zero, which implies SAP. Here, we see that we cannot reject SAP against alternative of general piglog (budget shares linear in $\ln x_h$) preferences.

We assumed piglog (budget shares linear in $\ln x_h$) preferences because they simplify our analysis relative to alternative specifications, and because, as documented earlier, they are widely used in practice, having been shown to often fit data well. Moreover, empirical violations of piglog are generally found at higher income levels (see, e.g., Banks, Blundell, and Lewbel 1997), which further suggests its adequacy in our application. To test the piglog assumption, we reestimate the model appending a quadratic term ($\ln x_h)^2$ to equation (9). The chi-square test statistic for the hypothesis that these quadratic terms can be excluded has 3 degrees of freedom. Its sample value given SAP is 6.6, with a p-value of 0.09, and its sample value relaxing SAP (but given DF) is 5.0, with a p-value of 0.17. We take this as evidence that the piglog model is acceptable.

Another potential objection to our specification is the assumption that multiple children share a utility function. We note that even this is an improvement over many empirical collective household applications, where only adults are assumed to have utility functions and expenditures on children are treated only as public goods consumed by adults. Nevertheless, we tried re-estimating the model using just the 881 households that have exactly one child. The results were similar in that the distribution factors were relevant given SAP and that SAP was not rejected. However, the distribution factors were now only marginally statistically significant in the DF model where the restriction of SAP was removed. We do not pursue this alternative further, due to its relatively small sample size.

Overall, we find that identification just based on distribution factors is feasible, but that it comes at a high cost in terms of precision of the resulting estimated resource shares. Second, we do not reject the SAP restrictions that DLP used to obtain identification. We therefore suggest that one should use both SAP and (when available) distribution factors to
obtain precisely estimated resource shares.

5.4 Deterministic Variation in Resource Shares

Before turning to random variation in resource shares, we first examine how these shares vary as functions of observed covariates. Let $\mu_{jh} = \mu_j(d_h, z_h)$. Table 3 summarizes the variation in $\mu_{jh}$, that is, the variation in estimated resource shares due only to variation in observed demographic variables $z$ and distribution factors $d$.

<table>
<thead>
<tr>
<th>Table 3: Estimated Deterministic Resource Shares</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAP</td>
</tr>
<tr>
<td>man all</td>
</tr>
<tr>
<td>d</td>
</tr>
<tr>
<td>s=1</td>
</tr>
<tr>
<td>s=2</td>
</tr>
<tr>
<td>s=3</td>
</tr>
<tr>
<td>s=4</td>
</tr>
<tr>
<td>woman all</td>
</tr>
<tr>
<td>d</td>
</tr>
<tr>
<td>s=1</td>
</tr>
<tr>
<td>s=2</td>
</tr>
<tr>
<td>s=3</td>
</tr>
<tr>
<td>s=4</td>
</tr>
<tr>
<td>children all</td>
</tr>
<tr>
<td>d</td>
</tr>
<tr>
<td>s=1</td>
</tr>
<tr>
<td>s=2</td>
</tr>
<tr>
<td>s=3</td>
</tr>
<tr>
<td>s=4</td>
</tr>
</tbody>
</table>

Descriptive statistics on estimated resource shares for 2887 households.


For men, women and children, we give the mean and standard deviations (across $z$ and $d$ values) of $\mu_{jh}$ in the row labeled “all”. The mean and standard deviation of $\mu_{jh}$ just due to variation in distribution factors is given in the row denoted “d”. For men, the average resource share is 0.3283, and the average contribution from observed distribution factors is $-0.0025$. The standard deviation of resource shares for men is 0.0782 and that of the
contribution from distribution factors is 0.0194, about a quarter as large. The relative size of variation induced by distribution factors is a bit smaller for women and children, where that variation accounts for roughly one sixth of overall variation in resource shares. The fact that resource shares determine 15 to 25 per cent of the variation in estimated resource shares lends further credibility to the relevance of these particular distribution factors, as required for identification of the DF model.

The lower block in each panel gives the average and standard deviation of resource shares for households with varying numbers of children. The pattern of children’s shares rising at the expense of adult shares conditional on demographic characteristics (observed in Table 2) remains evident here where we average across those demographic characteristics.

5.5 Estimates of Random Resource Shares

Here we provide estimates of the random variation in resources shares, based on the estimated models reported above. Recall that the resource shares $\eta_{jh}$ are given by $\eta_{jh} = \mu_{jh} U_{jh}/e_h$, where $\mu_{jh} = \mu_j (d_h, z_h)$ is the deterministic part of $\eta_{jh}$, while $U_{jh}/e_h$ is the idiosyncratic random component (with $e_h$ being the Berkson measurement errors).

In Table 4, we present, separately for men, women and children, summary statistics of the random resource shares $\eta_{jh}$, along with their deterministic components $\mu_{jh}$ (including the subcomponents that depend on distribution factors $d$), and their random components $U_{jh}$. As noted above, for this table, we drop the 5 households with $\mu_{jh} < 0.01$ for any $j$.

This table incorporates a number of checks on the model. First, while our estimates of $U_{jh}$ penalize the difference between the sample average $U_{jh}$ and one, we do not impose the constraint implied by the model that the mean of $U_{jh}$ exactly equal one. Nevertheless, our estimates come close to satisfying this constraint. As seen in Table 4, with SAP imposed the sample average values of $U_{jh}$ for men, women and children are 1.00, 1.03 and 1.01, respectively, while for DF, the sample averages are 1.00, 1.03 and 1.07, respectively.

In Table 4 we also report the correlations between the random resource share shifters, $U_{jh}$, and the deterministic part of resource shares, $\mu_{jh}$. Since $\mu_{jh}$ only varies as a function of exogenous covariates, we would expect the sample correlations between $\mu_{jh}$ and $U_{jh}$ to be near zero if the model is correctly specified. Imposing SAP, these correlations are $-0.03$, $0.04$ and $-0.07$, respectively, for men, women and children, while for DF estimates, these correlations are $-0.01$, $0.03$ and $-0.08$, respectively.

Another implication of the model that we do not impose is that the true mean of the Berkson errors $e$ equals one. Our estimated sample mean of $e$ is 1.03, which is very close to what the model predicts.
Now we turn to the distribution of resource shares $\eta_{jh}$. As we saw in Table 3, there is substantial variation in resource shares driven by observed covariates. In Table 4, we consider how much additional variation in resource shares emerges when we allow for possible unobserved distribution factors.

<table>
<thead>
<tr>
<th>Table 4: Random Resource Shares</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>man</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>woman</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>children</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Summary statistics on resource shares and $U_j$ for 2882 households.

Imposing SAP estimates, for men, the standard deviation of $\mu_{jh}$ (the deterministic component of $\eta_{jh}$) is roughly 8 percentage points, while the standard deviation of $\eta_{jh}$ is roughly 20 percentage points. So accounting for the random component of resource shares (and hence accounting for unobserved distribution factors) more than doubles the observed variation in resource shares.

The estimates for women and children show a similar pattern. For women, the standard deviation of $\mu_{jh}$ is 7 percentage points but that of $\eta_{jh}$ is 19 percentage points. For children, the standard deviation of $\mu_{jh}$ is 9 percentage points but that of $\eta_{jh}$ is 20 percentage points.

These random resource share estimates were obtained using a model that was identified by Theorem 1, based on distribution factors. The left-hand columns of Table 4 gives comparable...
values for the resource share distribution in the model that imposes SAP, and so has levels identified by DLP. Here, we see much the same pattern. Random resource shares have more than twice the standard deviation of resource shares that vary only with observed covariates.

The bottom line is that unobserved distribution factors appear to be at least as important as observed covariates in determining the variation in resource shares. These results suggest that searching for additional determinants of intra-household inequality (i.e., additional distribution factors) may be a fruitful area for future policy research.

### 5.6 Estimated Poverty Rates

In our data, 41 per cent of households have household total expenditures less than a standard household-level poverty line (defined as the sum of 2 US dollars per adult plus 1.20 US dollars per child per day). Households vary by size, and larger households are more likely than smaller ones to fall below this threshold. As a result, 43 per cent of people in our sample live in households with household total expenditures below the threshold. However, this calculated poverty rate ignores the fact that individuals within households may have unequal access to household resources. As noted in DLP, identification of the levels of resource shares allows us to calculate poverty rates at the person-level, rather than at the household-level, in a way that accounts for variation in access to household resources. In this subsection, we go beyond DLP to show the additional effect that accounting for random variation in resource shares has on the measurement of poverty at the individual level.

Table 5 gives estimated poverty rates for men, women and children using just the deterministic (fixed) part of resource shares, $\mu_{jh}$, and also gives poverty rates that account for the random component of resource shares, $\eta_{jh}$. To estimate the poverty rate for men, for example, for each household $h$ we compute the father’s estimated resources, equal to $\mu_{\text{man},h}$ times $x_h$ or $\eta_{\text{man},h}$ times $x_h$ (the former for fixed resource shares and the latter for random resource shares) and we count the fraction of households for which these estimated men’s resources are less than the poverty threshold. For children, the personal budget of each child in a household is equal to $\mu_{\text{children},h}$ times $x_h/s$ or $\eta_{\text{children},h}$ times $x_h/s$, where $s$ is the number of children in the household. The poverty thresholds we use are 2 US dollars per day for men and for women, and 1.20 US dollars per day (60% of the adult world poverty line, as implied by OECD standard equivalence scales) for children.
Table 5: Measured Poverty

<table>
<thead>
<tr>
<th></th>
<th>SAP</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fixed</td>
<td>random</td>
</tr>
<tr>
<td>person-level</td>
<td>men 37.8%</td>
<td>42.9%</td>
</tr>
<tr>
<td></td>
<td>women 38.2%</td>
<td>45.0%</td>
</tr>
<tr>
<td></td>
<td>children 58.8%</td>
<td>62.3%</td>
</tr>
</tbody>
</table>

Estimated poverty rates, from 2882 observations.


Similar to DLP, we find lower poverty rates for men than for women or for children. In Table 5 the estimated poverty rates based on SAP estimates and fixed deterministic resource shares $\mu_{jh}$ are 38 per cent for men and women, and 59 per cent for children. Estimated poverty rates based on DF estimates are 32 per cent for men, 38 per cent for women and 64 per cent for children. These rates are much lower than those reported in DLP, which was based on data from 12 years earlier, which similarly had much higher household level poverty rates.

We find that using $\eta_{jh}$, and thereby including random heterogeneity in resource shares, results in higher estimated poverty rates. Imposing SAP, our estimated poverty rates based on $\eta_{jh}$ are 43 per cent for men, 45 per cent for women, and 62 per cent for children. The estimated rates are 5 percentage points higher for men, 7 percentage points higher for women and for 5 percentage points higher for children, relative to estimates based on fixed resource shares $\mu_{jh}$. For DF estimates, the estimated poverty rates are 7 percentage points higher for men, 9 percentage points higher for women and unchanged for children.

The reason allowing for random variation in resource shares tends lower estimate poverty rates is that the poverty line is below the mode of the (unimodal) $\mu_{jh}x_h$ distribution for each $j$. The random component $U_{jh}$ moves some fraction of people who had $\mu_{jh}x_h$ near to, but above, the poverty line down below it, and pushes a different fraction of people who were below the line up above it. The poverty line is in the upward sloping part of the density of $\mu_{jh}x_h$, so there are more people close to but above the line than those close to but below the line. Hence more are randomly moved from above to below than are randomly moved from below to above. Consequently, accounting for random resource shares tends to increase measured poverty in our data.\(^{14}\)

We can summarize these results as follows. Calculating poverty rates in the standard way, at the household level, assumes equal sharing within the household and thereby ignores

\(^{14}\)A previous version of this paper used Malawian data from 2004, and found that random resource shares implied lower poverty rates. The reason is that, in 2004, roughly 90 per cent of Malawian households were poor based on the 2 dollar per adult measure, so that the poverty line was above the mode of the income distribution. Thus, accounting for random resource shares had the opposite effect in 2004 than it did 12 years later when “only” 40 per cent of households were poor.
actual intra-household allocation. Ignoring intra-household allocations overestimates men’s poverty rates, and underestimates the poverty rates of women and children, as reported by DLP and others (and reconfirmed here). However, in addition, failing to account for random variation in resource shares results in estimated poverty rates that are biased downwards for men, women and children in our data. Essentially, some individuals who seem non-poor on the basis of their observed characteristics (demographics and distribution factors) are in fact poor due to unobserved characteristics.

6 Conclusions

In this paper, we provide new resource share identification results. In Theorem 1 we show identification of the distribution of resource shares (around an unknown level) that can vary randomly across households. In Theorem 2 we show identification of the level of resource shares that does not require previously needed assumptions regarding similarity of tastes across households of varying compositions or across people within households. This second theorem uses a new way to exploit variation in so-called distribution factors, that is, variables that affect the allocation of resources within households but do not affect the tastes of individuals.

We apply our results to household survey data from Malawi. We find that identification using distribution factors is empirically practical, but can result in imprecise estimates if distribution factors are too highly correlated with other regressors. We use our results to test some of the behavioral restrictions that previous results required for identification. We find that the similarity across people (SAP) restriction proposed by Dunbar, Lewbel and Pendakur (2013) for identification is an acceptable restriction, and can be combined with our method of using observed distribution factors to yield more efficient resource share estimates.

We find that unobserved factors (i.e., random variation in resource shares) accounts for as much or more variation in resource shares as do observed covariates. Accounting for this random variation in resource shares across household members matters a great deal for poverty analyses. In particular, in our data accounting for such unobserved heterogeneity results in higher estimated poverty rates.
7 Bibliography

References


8 Appendix: Proofs and Additional Results

This Appendix has three sections. The first is derivation of the household demand equations for the private assignable goods. The second give proofs of our Theorems. The third describes an estimator of the resource share distribution for general demand equation functional forms, instead of the estimator given the text that was specific to the class of functional forms we use for our empirical analysis.

8.1 Derivation of household demand equations of private assignable goods

Here we summarize the derivation of equation (1), which is based on same machinery used in BCL and DLP to analyze collective household models where most goods can be shared to varying extents. Unlike DLP, we explicitly include price variation, unlike BCL, we focus on assignable goods and include more than two household members, and unlike both DLP and BCL we explicitly include distribution factors, both observable $d$ and unobservable $v$.

Start by assuming the household chooses what to consume using the program

$$\max_{y_1, \ldots, y_J, r_1, \ldots, r_J} \tilde{V} \left[ V_1 (y_1, r_1, z), \ldots, V_J (y_J, r_J, z) \mid d, z, p/x^*, v \right]$$

such that

$$r = \sum_{j=1}^{J} r_j$$

and

$$x^* = p'_A(z) r + \sum_{j=1}^{J} p_j y_j$$

where $V_j (y_j, r_j, z)$ for $j = 1, \ldots, J$ is the utility function of household member $j$, and the function $\tilde{V}$ describes the social welfare function or bargaining process of the household, which exists because the household is pareto efficient.

Recall $z$ denotes a vector of observable attributes of households and their members like age, education, and number of children. Household attributes $z$ may affect preferences, and so appear inside the utility functions $V_j$. These $z$ variables may also affect the bargaining process or social welfare function given by $\tilde{V}$ (by, e.g., affecting the relative bargaining power of members), and as a result may affect resource shares.


We have scalars \( y_1, \ldots, y_J \) that are the quantities of private, assignable goods, where member \( j \) has quantity \( y_j \) in his or her utility function, and does not have \( y_\ell \) for all \( \ell \neq j \) in his or her utility function. Each member’s utility function also depends on a quantity vector of other goods \( r_j \). The market prices of these goods are given by the vector \( p_r \). The square matrix \( A(z) \) is what BCL call a linear consumption technology function over goods. Having \( A(z) \) differ from the identity matrix is what allows goods in \( r \) to be partly shared and/or consumed jointly. In particular, \( A(z) r \) equals the quantity vector of these goods that the household actually purchases, while \( r = \sum_{j=1}^{J} r_j \) is total quantity vector of these goods that the household consumes. These quantities are not the same due to sharing and joint consumption. The smaller an element of \( A(z) r \) is relative to the corresponding element of \( r \), the more that good is shared or jointly consumed. See BCL for details. The vector of all prices \( p \) includes, \( p_r \), the vector of prices of the elements of \( r \), and \( p_1, \ldots, p_J \), the prices of the private assignable goods \( y_1, \ldots, y_J \).

What makes the vectors \( d \) and \( v \) be distribution factors (observed and unobserved, respectively) in the model is that they appear only as arguments of \( \tilde{V} \), and so only affect the allocation of resources within the household, but not the tastes of the individual household members.

Applying duality theory and decentralization welfare theorems, it follows from BCL that the household’s program above is equivalent to a program where each household member \( j \) chooses what to consume using the program

\[
\max_{y_j, r_j} V_j(y_j, r_j, z) \quad \text{such that} \quad \eta_j(p, x^*, z, d, v)x^* = p_r A(z) r_j + p_j y_j
\]

where \( \eta_j = \eta_j(p, x^*, z, d, v) \) is the resource share of member \( j \), that is, \( \eta_j \) is the fraction of total household resources \( x^* \) that are allocated to member \( j \). This member then chooses quantities \( y_j \) and the vector \( r_j \) subject to a linear budget constraint. The vector \( p_r A(z) \) equals the vector of shadow prices of goods \( r \). These shadow prices for the household are lower than market prices, due to sharing. Being private and assignable, the shadow price of each \( y_j \) equals its market price \( p_j \). The shadow budget for member \( j \) is \( \tilde{x}_j = \eta_j x^* \). As shown in BCL, the resource share functions \( \eta_j(p, x^*, z, d, v) \) for each member \( j \) in general depend on the function \( \tilde{V} \) and on the utility functions \( V_1, \ldots, V_J \).

BCL show that the more bargaining power a household member has (i.e., the greater is the weight of his or her utility function in \( \tilde{V} \)), the larger is their resource share \( \eta_j \). Resource shares \( \eta_j \) all lie between zero and one, and resource shares sum to one, that is, \( \sum_{j=1}^{J} \eta_j = 1 \).

As in DLP, we will not work with the household demand functions of all goods (which, as shown in BCL, can be rather complicated). Instead, we only make use of the demand
functions of the private assignable goods $y_j$, which are simpler. Since equation (12) is an ordinary utility function maximised under a linear budget constraint (linear in shadow prices and a shadow budget), the solution to equation (12) is a set of Marshallian demand equations for $y_j$ and $r_j$. Let $h_j(\tilde{x}_j, p, z)$ be the Marshallian demand function of person $j$ for their private assignable good, that is, $h_j(\tilde{x}_j, p, z)$ is the quantity person $j$ in a household with member attributes $z$ would demand of their assignable good if they had a budget equal to their shadow budget $\tilde{x}_j$ and faced the within-household shadow price vector that corresponds to the market price vector $p$. Since each $y_j$ is private and assignable, the quantity $y_j$ that member $j$ chooses to consume equals the quantity of this good that the household buys. It therefore follows from the above that the household’s quantity demand of each private assignable good $y_j$ is given by equation (1).

A few general comments regarding BCL (which then also apply to DLP and the present paper) are worth noting. First, consider public goods. In collective models, public goods are assumed to be completely shared, with the quantity the household purchases of each public good appearing directly in each member’s utility function. In BCL, goods are instead partly shared, or partly jointly consumed. In particular, elements of the $A(z)$ matrix describe how much each good is shared. The lower is an element on the diagonal of $A(z)$, the greater is the extent to which that good is shared, and hence the more like a public good that good is. As a result, our estimated resource shares incorporate the extent to which goods are partly public.

Second, consider externalities within the household. BCL does not permit externalities wherein one member get utility from the consumption of individual goods of other members. However, BCL does permit externalities in the form of caring preferences, that is, the utility level of one household member can depend on the attained utility levels of other members. This does not affect the estimation of resource shares (and they still represent the consumption of each member). But if caring preferences are present, then the link between resources shares and household bargaining power can be affected, e.g., someone with high power and substantial caring preferences could end up with a low resource share.

Finally, note that DLP do not identify $A(z)$ or indifference scales, due to the data limitations that DLP copes with relative to BCL. Since the present paper builds on DLP, it is an open question whether $A(z)$ or indifference scales could be identified using some version of the methods provided here.
8.2 Proofs

PROOF OF THEOREM 1: Let $h_j^{-1}(y_j, p, z)$ denote the inverse of the function $h_j$ with respect to its first argument, which exists by Assumption A1. Let $F_{Y_j}(y_j | p, x, d, z)$ be the identified conditional distribution of $Y_j$. Let $\tilde{\eta}_j = c\eta_j$ and let $F_{\tilde{\eta}_j}(\tilde{\eta}_j | p, d, z)$ be the unknown conditional distribution of $\tilde{\eta}_j$. Then for any $y_j \in \text{supp}(Y_j | p, x, d, z)$,

$$F_{Y_j}(y_j | p, x, d, z) = \Pr(h_j(\tilde{\eta}_j x, p, z) \leq y_j | p, x, d, z) = \Pr\left(\tilde{\eta}_j \leq \frac{h_j^{-1}(y_j, p, z)}{x} \Big| p, x, d, z\right)$$

$$= F_{\tilde{\eta}_j}\left(\frac{h_j^{-1}(y_j, p, z)}{x} \Big| p, x, d, z\right) = F_{\tilde{\eta}_j}\left(h_j^{-1}(y_j, p, z) \Big| p, d, z\right)$$

where the last equality follows from Assumption A3. By continuity of the distribution of $\tilde{\eta}_j$, the distribution function $F_{\tilde{\eta}_j}$ is differentiable, and

$$\frac{-x\partial F_{Y_j}(y_j | p, x, d, z) / \partial y_j}{\partial F_{Y_j}(y_j | p, x, d, z) / \partial x} = \frac{-x\partial h_j^{-1}(y_j, p, z) / \partial y_j}{\partial h_j^{-1}(y_j, p, z) / \partial x} = \frac{\partial \ln h_j^{-1}(y_j, p, z)}{\partial y_j}$$

It follows by the continuity assumptions and interval support of $x$ that $\text{supp}(Y_j | p, z)$ is an interval. Let $\kappa_j(p, z)$ denote any given nonzero element in this support. Define the identified function $r_j$ by

$$r_j(y_j, p, z) = \exp\left(\int_{\kappa_j(p, z)}^{y_j} E\left(\frac{-x\partial F_{Y_j}(y_j | p, x) / \partial y_j}{\partial F_{Y_j}(y_j | p, x) / \partial x} \mid y_j, p, z\right) dy_j\right)$$

$$= \exp\left(\int_{\kappa_j(p, z)}^{y_j} \frac{\partial \ln h_j^{-1}(y_j, p, z)}{\partial y_j} dy_j\right)$$

$$= h_j^{-1}(y_j, p, z) c_j(p, z)$$

where $c_j(p, z) = 1/h_j^{-1}(\kappa_j(p, z), p, z)$ is an unknown function.

Now $h_j^{-1}(Y_j, p, z) = \tilde{\eta}_j x$ so $r_j(Y_j, p, z) / x = \tilde{\eta}_j c_j(p, z)$. Applying the same derivation used above to relate $F_{Y_j}$ to $F_{\tilde{\eta}_j}$ we have

$$F_{Y_1, \ldots, Y_J}(y_1, \ldots, y_J | p, x, d, z) = F_{\tilde{\eta}_1, \ldots, \tilde{\eta}_J}\left(\frac{h_1^{-1}(y_1, p, z)}{x}, \ldots, \frac{h_J^{-1}(y_J, p, z)}{x} \Big| p, d, z\right)$$

$$= F_{\tilde{\eta}_1, \ldots, \tilde{\eta}_J}\left(\frac{r_1(y_1, p, z)}{c_1(p, z) x}, \ldots, \frac{r_J(y_J, p, z)}{c_J(p, z) x} \Big| p, d, z\right)$$

Since the $F_{Y_1, \ldots, Y_J}$ is known and the $r_j(Y_j, p, z)$ functions are identified, this shows that the $F_{\tilde{\eta}_1, \ldots, \tilde{\eta}_J}$ distribution is identified up to the unknown location terms $c_1(p, z), \ldots, c_J(p, z)$.
Equivalently, we can say that the joint distribution of \( r_1(Y_j, p, z) / x, \ldots, r_J(Y_j, p, z) / x \) conditional on \( p, d, z \) is identified from identification of \( F_Y(Y_1, \ldots, Y_j | p, x, d, z) \) and of the functions \( r_j(y_j, p, z) \). This joint distribution of \( r_1(Y_j, p, z) / x, \ldots, r_J(Y_j, p, z) / x \) conditional on \( p, d, z \) equals the joint distribution of \( \tilde{\eta}_1 c_1(p, z), \ldots, \tilde{\eta}_J c_J(p, z) \) conditional on \( p, d, z \), which proves the Theorem.

PROOF OF THEOREM 2: Define the identified function \( t_j \) by

\[
\begin{align*}
t_j(p, d, z) &= E\left( \frac{r_j(Y_j, p, z)}{x} \bigg| p, d, z \right) = E(\eta_j c_j(p, z) | p, d, z) \\
&= E(e) E(\eta_j | p, d, z) c_j(p, z)
\end{align*}
\]

so

\[
\frac{t_j(p, d_k, z)}{t_j(p, d_1, z)} E(\eta_j | p, d_1, z) = E(\eta_j | p, d_k, z)
\]

It follows that \( T(p, z) \) defined in Assumption A4 equals the \( J \) by \( J \) matrix that has \( t_j(p, d_k, z) / t_j(p, d_1, z) \) in the row \( k \) and column \( j \) position, and therefore that the matrix \( T(p, z) \) is identified. We also have

\[
\sum_{j=1}^{J} \frac{t_j(p, d_k, z)}{t_j(p, d_1, z)} E(\eta_j | p, d_1, z) = \sum_{j=1}^{J} E(\eta_j | p, d_k, z) = E\left( \sum_{j=1}^{J} \eta_j | p, d_k, z \right) = 1
\]

Let \( E(\eta | p, d_1, z) \) be the vector of elements \( E(\eta_j | p, d_1, z) \) for \( j = 1, \ldots, J \), and let \( 1_J \) denote the \( J \) vector of ones. Then the above equation for \( k = 1, \ldots, J \) is equivalent to \( T(p, z) E(\eta | p, d_1, z) = 1_J \), and so, using Assumption A4, \( E(\eta | p, d_1, z) = T(p, z)^{-1} 1_J \), and therefore \( E(\eta_j | p, d_1, z) \) is identified for \( j = 1, \ldots, J \). We can then identify the functions \( c_j(p, z) \) for \( j = 1, \ldots, J \) by \( c_j(p, z) = t_j(p, d_1, z) / E(\eta_j | p, d_1, z) \). By Theorem 1, the joint distribution of \( \tilde{\eta}_1 c_1(p, z), \ldots, \tilde{\eta}_J c_J(p, z) \) conditional on \( p, d, z \), is identified, and since \( c_j(p, z) \) for \( j = 1, \ldots, J \) is identified, we can divide elements of one by the other to conclude that the joint distribution of \( \tilde{\eta}_1, \ldots, \tilde{\eta}_J \) conditional on \( p, d, z \), is identified. Now \( e = \tilde{\eta}_1, \ldots, \tilde{\eta}_J \) and therefore the distribution of \( e \) is identified. We have that the joint distribution of \( \ln \tilde{\eta} = \ln \eta + \ln e \) conditional on \( p, d, z \), is identified, and the distribution of \( \ln e \) is identified and independent of \( \ln \eta \), so by a deconvolution the joint distribution of \( \ln \eta \) and hence of \( \eta \), conditional on \( p, d, z \), is identified.

### 8.3 A General Estimator

In the text we provided an estimator for the model that depended in part on features of our chosen piglog functional form for Engel curves. Here we describe a more general estimation
method that doesn’t depend on specific properties of the piglog, but is more complicated to apply in that it requires estimating an inverted demand system. Suppress $z$ for the moment: both $h_j$ and $\eta_j$ depend on it. Suppress $p$ also and for now consider an Engel curve exercise. Assume we have some parameterization of the clothing demand functions $h_j$, so the model is

$$Y_j = h_j(x\eta_j e, z \mid \theta_j)$$

where $\theta_j$ is a vector of clothing engel curve parameters, $\eta_j$ are random resource shares, and $e$ is multiplicative measurement error on total expenditures.

For $j = 1, 2, 3$ assume that $E(\eta_j \mid x, d, z) = E(\eta_j \mid d, z)$, so on average resource shares do not depend on $x$. Let

$$s_j (d, z \mid \gamma) = E(\eta_j \mid d, z)$$

so $s_j$ is an assumed parameterised functional form for mean resource shares, which depends on a parameter vector $\gamma$. Resource shares sum to one, that is, $\eta_1 + \eta_2 + \eta_3 = 1$, so $s_1 (d, z \mid \gamma_1) + s_2 (d, z \mid \gamma_2) + s_3 (d, z \mid \gamma_3) = 1$.

We assume $e$ is multiplicative measurement error, with $E[e \mid d, x, z, \eta_1, \eta_2, \eta_3] = 1$. Let $Y_j$ be normal goods, so that $h_j$ is invertible. Let $g_j(.) = h_j^{-1}(.)$ be the inverse demand functions for each $j$. Then by definition

$$0 = g_j(Y_j \mid \theta_j) - x\eta_j e$$

Taking conditional expectations gives

$$0 = E[g_j(Y_j \mid \theta_j) - x\eta_j e \mid d, x, z, \eta_1, \eta_2, \eta_3]$$
$$= E[g_j(Y_j \mid \theta_j) \mid d, x, z, \eta_1, \eta_2, \eta_3] - x\eta_j E[e \mid d, x, z, \eta_1, \eta_2, \eta_3]$$
$$= E[g_j(Y_j \mid \theta_j) \mid d, x, z, \eta_1, \eta_2, \eta_3] - x\eta_j$$

Now take conditional expectations of this, just conditioning on $d, x, z$, and applying the law of iterated expectations to get

$$0 = E[g_j(Y_j \mid \theta_j) \mid d, x, z] - xE[\eta_j \mid d, x, z]$$
$$= E[g_j(Y_j \mid \theta_j) \mid d, x, z] - xE[\eta_j \mid d, z]$$
$$= E[g_j(Y_j \mid \theta_j) \mid d, x, z] - xs_j (d, z \mid \gamma)$$
$$= E[g_j(Y_j \mid \theta_j) - xs_j (d, z \mid \gamma) \mid d, x, z]$$

Based on this last equation estimate the parameters $\theta_j$ and $\gamma$ for $j = 1, 2, 3$ by GMM, using
unconditional moments of the form

\[
0 = E \left[ \left[ g_j(Y_j | \theta_j) - x s_j (d, z | \gamma) \right] r (d, x, z) \right] = 0
\]

for \( j = 1, 2, 3 \), where instruments \( r (x, d, z) \) are functions of the variables \( x, d, \) and \( z \) that we choose, and \( g_j \) and \( s_j \) have the functional forms we choose.

Now

\[
g_1(Y_1 | \theta_1) + g_2(Y_2 | \theta_2) + g_3(Y_3 | \theta_3) = (\eta_1 + \eta_2 + \eta_3) e x = e x
\]

So, given estimates of the parameters \( \hat{\theta}_j \), we can recover estimates of the value of \( e, \eta_1, \eta_2, \) and \( \eta_3 \) for each consumer \( j \) in each household \( i \) by

\[
\hat{e}_i = \frac{g_1(Y_{1i} | \hat{\theta}_j) + g_2(Y_{2i} | \hat{\theta}_2) + g_3(Y_{3i} | \hat{\theta}_3)}{x_i}
\]

and

\[
\hat{\eta}_{ji} = \frac{g_j(Y_{ji} | \theta_j)}{\hat{e}_i x_i}
\]

We could estimate the population distribution of \( e \) and the joint distribution of resource shares by kernel density estimation using the above estimates of each consumer’s \( \hat{e}_i, \) and \( \hat{\eta}_{ji} \) as data. Also, given \( \hat{e}_i, \) and \( \hat{\eta}_{ji}, \) the demand function of a consumer \( j \) in household \( i \) is given by

\[
h_j(\hat{e}_i \hat{\eta}_{ji} x_i | \hat{\theta}_j).
\]

9 Appendix Tables

Table A1 gives complete estimates of our SAP and DF models. These are unformatted Stata output.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimation</th>
<th>Std Err</th>
<th>Parameter</th>
<th>Estimation</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \eta_m )</td>
<td>0.3145</td>
<td>0.0428</td>
<td>( \gamma_m )</td>
<td>0.0546</td>
<td>0.0119</td>
</tr>
<tr>
<td>c2</td>
<td>-0.0255</td>
<td>0.0345</td>
<td>c2</td>
<td>-0.0033</td>
<td>0.0086</td>
</tr>
<tr>
<td>c3</td>
<td>-0.1001</td>
<td>0.0410</td>
<td>c3</td>
<td>-0.0051</td>
<td>0.0115</td>
</tr>
<tr>
<td>c4</td>
<td>-0.0804</td>
<td>0.0527</td>
<td>c4</td>
<td>0.0200</td>
<td>0.0156</td>
</tr>
<tr>
<td>z1</td>
<td>0.0755</td>
<td>0.0289</td>
<td>z1</td>
<td>0.0068</td>
<td>0.0072</td>
</tr>
<tr>
<td>z2</td>
<td>-0.0020</td>
<td>0.0369</td>
<td>z2</td>
<td>0.0492</td>
<td>0.0125</td>
</tr>
<tr>
<td>z3</td>
<td>0.0092</td>
<td>0.0297</td>
<td>z3</td>
<td>0.0028</td>
<td>0.0026</td>
</tr>
<tr>
<td>z4</td>
<td>0.0046</td>
<td>0.0082</td>
<td>z4</td>
<td>0.0030</td>
<td>0.0024</td>
</tr>
<tr>
<td>z7</td>
<td>0.0627</td>
<td>0.0305</td>
<td>z7</td>
<td>0.0053</td>
<td>0.0082</td>
</tr>
<tr>
<td>z8h</td>
<td>-0.0375</td>
<td>0.0192</td>
<td>z8h</td>
<td>0.0161</td>
<td>0.0234</td>
</tr>
<tr>
<td>z8s</td>
<td>-0.0832</td>
<td>0.0393</td>
<td>z8s</td>
<td>0.0017</td>
<td>0.0049</td>
</tr>
<tr>
<td>z9h</td>
<td>0.0226</td>
<td>0.0181</td>
<td>z9h</td>
<td>0.0047</td>
<td>0.0076</td>
</tr>
<tr>
<td>z15</td>
<td>0.0017</td>
<td>0.0074</td>
<td>z15</td>
<td>0.0038</td>
<td>0.0027</td>
</tr>
<tr>
<td>z16</td>
<td>-0.1229</td>
<td>0.0547</td>
<td>z16</td>
<td>0.0038</td>
<td>0.0025</td>
</tr>
<tr>
<td>z18</td>
<td>0.0276</td>
<td>0.0379</td>
<td>z18</td>
<td>0.0049</td>
<td>0.0026</td>
</tr>
<tr>
<td>df4</td>
<td>0.0008</td>
<td>0.0141</td>
<td>df4</td>
<td>0.0001</td>
<td>0.0015</td>
</tr>
<tr>
<td>z2</td>
<td>-0.0010</td>
<td>0.0012</td>
<td>z2</td>
<td>0.0037</td>
<td>0.0081</td>
</tr>
<tr>
<td>z6</td>
<td>-0.0084</td>
<td>0.0050</td>
<td>z6</td>
<td>-0.0094</td>
<td>0.0094</td>
</tr>
<tr>
<td>z7</td>
<td>-0.0067</td>
<td>0.0142</td>
<td>z7</td>
<td>-0.0002</td>
<td>0.0014</td>
</tr>
<tr>
<td>z8h</td>
<td>-0.0049</td>
<td>0.0151</td>
<td>z8h</td>
<td>0.0000</td>
<td>0.0015</td>
</tr>
<tr>
<td>z9h</td>
<td>0.0006</td>
<td>0.0201</td>
<td>z9h</td>
<td>0.0000</td>
<td>0.0011</td>
</tr>
<tr>
<td>z15</td>
<td>0.0016</td>
<td>0.0035</td>
<td>z15</td>
<td>0.0001</td>
<td>0.0009</td>
</tr>
<tr>
<td>z16</td>
<td>0.0003</td>
<td>0.0015</td>
<td>z16</td>
<td>0.0008</td>
<td>0.0004</td>
</tr>
<tr>
<td>z18</td>
<td>-0.0014</td>
<td>0.0013</td>
<td>z18</td>
<td>-0.0007</td>
<td>0.0003</td>
</tr>
<tr>
<td>z2</td>
<td>0.0059</td>
<td>0.0078</td>
<td>z2</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>z19</td>
<td>0.0164</td>
<td>0.0236</td>
<td>z19</td>
<td>0.0055</td>
<td>0.0001</td>
</tr>
<tr>
<td>f1</td>
<td>0.3834</td>
<td>0.3851</td>
<td>f1</td>
<td>0.0031</td>
<td>0.0060</td>
</tr>
<tr>
<td>f2</td>
<td>-0.0313</td>
<td>0.0343</td>
<td>f2</td>
<td>-0.0097</td>
<td>0.0076</td>
</tr>
<tr>
<td>c3</td>
<td>0.0036</td>
<td>0.0479</td>
<td>c3</td>
<td>0.0015</td>
<td>0.0033</td>
</tr>
<tr>
<td>c4</td>
<td>0.0027</td>
<td>0.0017</td>
<td>c4</td>
<td>0.0005</td>
<td>0.0000</td>
</tr>
<tr>
<td>c5</td>
<td>-0.0943</td>
<td>0.0305</td>
<td>c5</td>
<td>0.0005</td>
<td>0.0014</td>
</tr>
<tr>
<td>z2</td>
<td>0.0133</td>
<td>0.0401</td>
<td>z2</td>
<td>0.0005</td>
<td>0.0011</td>
</tr>
<tr>
<td>z5</td>
<td>0.0010</td>
<td>0.0101</td>
<td>z5</td>
<td>0.0036</td>
<td>0.0041</td>
</tr>
<tr>
<td>z6</td>
<td>0.0093</td>
<td>0.0092</td>
<td>z6</td>
<td>0.0045</td>
<td>0.0130</td>
</tr>
<tr>
<td>z7</td>
<td>0.1856</td>
<td>0.0948</td>
<td>z7</td>
<td>0.0103</td>
<td>0.0146</td>
</tr>
<tr>
<td>z8h</td>
<td>0.0210</td>
<td>0.0176</td>
<td>z8h</td>
<td>0.0002</td>
<td>0.0019</td>
</tr>
<tr>
<td>z9h</td>
<td>0.0682</td>
<td>0.0385</td>
<td>z9h</td>
<td>0.0002</td>
<td>0.0019</td>
</tr>
<tr>
<td>z10</td>
<td>0.0008</td>
<td>0.0095</td>
<td>z10</td>
<td>0.0019</td>
<td>0.0288</td>
</tr>
<tr>
<td>z11</td>
<td>0.0384</td>
<td>0.0504</td>
<td>z11</td>
<td>0.0109</td>
<td>0.0094</td>
</tr>
<tr>
<td>z12</td>
<td>-0.0947</td>
<td>0.1225</td>
<td>z12</td>
<td>0.0016</td>
<td>0.0077</td>
</tr>
<tr>
<td>d1</td>
<td>0.0055</td>
<td>0.0130</td>
<td>d1</td>
<td>0.0047</td>
<td>0.0104</td>
</tr>
<tr>
<td>d4</td>
<td>-0.0211</td>
<td>0.0111</td>
<td>d4</td>
<td>0.0044</td>
<td>0.0124</td>
</tr>
<tr>
<td>c1</td>
<td>0.0464</td>
<td>0.0078</td>
<td>c1</td>
<td>0.0017</td>
<td>0.0038</td>
</tr>
<tr>
<td>c2</td>
<td>-0.0948</td>
<td>0.0606</td>
<td>c2</td>
<td>-0.0015</td>
<td>0.0102</td>
</tr>
<tr>
<td>c3</td>
<td>-0.0148</td>
<td>0.0072</td>
<td>c3</td>
<td>0.0002</td>
<td>0.0025</td>
</tr>
<tr>
<td>c4</td>
<td>0.0083</td>
<td>0.0043</td>
<td>c4</td>
<td>0.0009</td>
<td>0.0011</td>
</tr>
<tr>
<td>c5</td>
<td>0.0025</td>
<td>0.0066</td>
<td>c5</td>
<td>0.0003</td>
<td>0.0017</td>
</tr>
<tr>
<td>c6</td>
<td>0.0023</td>
<td>0.0015</td>
<td>c6</td>
<td>0.0182</td>
<td>0.0211</td>
</tr>
<tr>
<td>c7</td>
<td>-0.0021</td>
<td>0.0013</td>
<td>c7</td>
<td>0.0007</td>
<td>0.0025</td>
</tr>
<tr>
<td>c8h</td>
<td>0.0002</td>
<td>0.0051</td>
<td>c8h</td>
<td>0.0121</td>
<td>0.0165</td>
</tr>
<tr>
<td>c8s</td>
<td>-0.0254</td>
<td>0.0148</td>
<td>c8s</td>
<td>-0.0030</td>
<td>0.0019</td>
</tr>
<tr>
<td>c9h</td>
<td>-0.0065</td>
<td>0.0159</td>
<td>c9h</td>
<td>0.0004</td>
<td>0.0016</td>
</tr>
<tr>
<td>c9s</td>
<td>-0.0035</td>
<td>0.0020</td>
<td>c9s</td>
<td>0.0000</td>
<td>0.0007</td>
</tr>
<tr>
<td>c10</td>
<td>0.0041</td>
<td>0.0023</td>
<td>c10</td>
<td>-0.0211</td>
<td>0.0165</td>
</tr>
<tr>
<td>c11</td>
<td>0.0005</td>
<td>0.0017</td>
<td>c11</td>
<td>-0.0014</td>
<td>0.0018</td>
</tr>
<tr>
<td>c12</td>
<td>0.0043</td>
<td>0.0066</td>
<td>c12</td>
<td>-0.0046</td>
<td>0.0077</td>
</tr>
<tr>
<td>c13</td>
<td>0.0093</td>
<td>0.0177</td>
<td>c13</td>
<td>0.0000</td>
<td>0.0035</td>
</tr>
</tbody>
</table>