An Engel Curve for Variety

Nicholas Li

University of Toronto

Abstract: I examine the source and welfare implications of differences in household consumption diversity. I document the existence of a positive correlation between household variety and expenditure to motivate a simple framework where households purchase more varieties to counteract diminishing returns to quantity but face location-specific costs of accessing variety. Estimating the model with Indian household data, I find that the increase in dietary diversity between 1983 and 2009 was mostly due to lower costs of accessing variety that resulted in large welfare gains. Urban households also benefit from a lower cost of accessing varieties than rural households. JEL codes: D12, F15, O12, R13

Acknowledgments: Thanks to Chang-Tai Hsieh, Pierre-Olivier Gourinchas, Ted Miguel, Yuriy Gorodnichenko, Mu-Jeung Yang, Xiaodong Zhu, Gustavo Bobonis, Rob McMillan, and seminar participants at UC Berkeley, Stanford, Chicago, London School of Economics, the World Bank, University of Toronto, University of Sydney and University of New South Wales, the editor and referees for comments. Financial support was provided by SSHRC, the Institute for Business and Economics Research and the Center for Equitable Growth at UC Berkeley. Contact: nick.li@utoronto.ca

1. Introduction

According to India's National Sample Survey, the average Indian household consumed 24 food varieties in 1983 but by 2009 this had risen to 36. In 2009 the average urban household in India consumed three more food varieties than the average rural household. The transition from a monotonous, staple-heavy diet to a diverse one seems ubiquitous to economic development and has been studied extensively from a nutrition perspective, but it has received little attention from consumer theory. There is a large empirical literature has quantified the magnitude and consumer welfare implications of market-level variety differences (Feenstra (1994), Broda and Weinstein (2006), Broda and Weinstein (2010), Handbury and Weinstein (2015), Handbury (2013), Couture (2015), Hsieh et al. (2016)), but differences in household-level variety may also be informative about consumer welfare. I address this gap in the literature by presenting several facts about household consumption diversity, a simple model consistent with these facts, and an application of the model to estimate the sources and welfare implications of differences in household food variety in India.

A major feature of the data is the positive association between household variety and household expenditure within a location, which I call a "variety Engel curve." This fact highlights both a limited demand for household variety, since poor households do not consume as many varieties as their richer neighbors, and a potential role for income differences in explaining household variety even in the absence of differences in the retail environment. I show that in India these variety Engel curves are shifted up in urban (relative to rural) locations and over time and provide evidence that both retail density and road access support the consumption of more varieties by households. Locations that facilitate the consumption of more varieties may thus benefit households through a household variety channel. I also show that households who consume more varieties – whether due to their income or retail environment – diversify their consumption by adding varieties that are marginal along multiple dimensions and shop more.

I develop a simple model consistent with these facts. Households with diminishing marginal utility from quantity (per variety) incur a variety cost that depends on their retail environment. High spending households optimally choose to consume more varieties and incur more variety costs. Locations where marginal varieties are relatively important (e.g. due to lower price or higher taste) or where the marginal cost of accessing variety is lower have higher household variety conditional on expenditure. The model generates log-linear variety Engel curves similar to the data and provides a decomposition of variety differences between any two households into an expenditure component, a variety marginality component, and a component related to the cost of accessing variety. The model's parameters can be used to estimate cost-of-living differences across locations that vary with income due to non-homotheticity of variety demand.

I estimate the model using Indian data on grain and vegetable consumption to quantify the sources and welfare implications of higher variety over time and in urban areas. The analysis reveals that differences in expenditures and the relative importance of marginal varieties play an important role in generating differences in variety, but that differences in variety costs explain most of the differences in variety across locations. The implied welfare gains from lower variety cost parameters are quite large over time (about 10% of expenditures for grains and 20% for vegetables on average) but more modest for urban versus rural locations (about 4%). These welfare gains are mostly biased towards higher income households with the exception of improved vegetable variety over time which has benefited poor households and more remote regions the most.

The paper makes two main contributions. The first is to provide an empirical analysis of household variety in a developing country setting. Most of the literature on consumption variety analyzes market-level differences in developed countries, or more recently middle-income countries (Atkin et al. (2016)). The welfare gains from these variety differences are usually interpreted as arising from a better match between the set of available varieties and the heterogeneous tastes of consumers with discrete choice preferences

(Anderson et al. (1992)). The distinct sources and implications of household variety when households have diminishing returns to quantity have received little attention. I show that while market and household variety can be correlated, household variety is shaped by the interaction of household income and the retail environment in subtle ways, such that market variety can be a poor predictor of household variety. Because shopping is an important input into household variety, differences in retail convenience and the willingness of richer households to travel farther to access more exotic varieties make the standard assumptions for market-level variety analysis – that all households in a given location have either costless or zero access to each variety – questionable.[1] In a setting like mine, with low income, low retail productivity, and high transaction costs (Lagakos (2015), Banerjee and Duflo (2011)), the variety of basic foods consumed by households is an important consumption margin and I provide a detailed analysis of this margin.

The second contribution is to provide a model for analyzing the source and welfare implications of household variety differences using only household consumption data. Models that allow households to purchase multiple varieties (Hendel (1999), Dube (2004), Wales and Woodland (1983), and Kim et al. (2002)) have been estimated by first observing a retailer choice set and then estimating the parameters that determine household variety. These models can allow for more flexible income and substitution effects and some allow changes in the choice set to affect household welfare through both diminishing returns to quantity and heterogeneous taste channels. However, the data on shopping patterns, retailer locations and assortments necessary to implement them are rarely available, making their application outside of the usual marketing context less feasible. Instead of using prior knowledge of the choice set to estimate the parameters that affect household variety, my approach uses knowledge of household consumption choices to estimate variety cost parameters that capture the (unobserved) differences in the retail environment relevant for household variety and welfare. My use of consumption data to estimate differences

[1] See Bronnenberg (2015) and Allender et al. (2013) for related contributions.

in the cost-of-living has similarities to approaches using food Engel curves (Hamilton (2001), Almas (2012)) or quality Engel curves (Bils and Klenow (2001a)) to measure price index bias, and is appealing in similar contexts where direct measurement is infeasible.

Section 2 presents four facts about variety Engel curves using data from India and other countries. Section 3 considers a model motivated by these facts and its relation to existing frameworks. Section 4 discusses model estimation and presents results on the source and welfare implications of variety differences in India, and section 5 concludes.

2. Descriptive facts

My analysis focuses on data from India's National Sample Survey (NSS), a typical developing country consumption survey. Households are asked to recall expenditures and quantities consumed during the previous 30 day period from a list of items. Defining a "variety" as the most disaggregated item recorded consistently between 1983 (38th round) and 2009-2010 (66th round), there are 134 total food items, 18 grain items and 29 vegetable items (see the bottom of Table 1 for a complete list of food varieties for these food categories). Table 1 presents summary statistics for mean variety per household over time and for rural and urban households. Consumption variety rose 50% for the average household between 1983 and 2009 and almost doubled for vegetables. Urban households consume higher variety than rural households. These mean variety differences sometimes, but not always, coincide with real expenditure differences. [2]

To better understand what underlies these patterns, I document four facts using the NSS and other Indian data that motivate the model developed in the next section. I also refer to supplemental results in the Appendix that use data from other countries to ad-

[2] I use the median unit values contained in the survey to calculate a Tornqvist price index across sector/years for each aggregation (food, grains, vegetables). I have also adjusted unit values for quality effects following the procedure in Deaton (1988) but this has little impact on the results. Home produced goods have their value imputed using farm-gate prices, while gifts and in-kind payments use local retail prices.

dress limitations of the Indian data and show that these facts are not specific to India.

A. Fact 1: Household variety increases with expenditure within location

The data group households by state, region, district (for some years), and first-stage sampling units of 10 households drawn from an anonymized rural village or urban block. Within any geographic aggregation, there is a strong (and approximately log-linear) positive relationship between household expenditure and household variety. This correlation is what I call an "Engel curve for variety." The top three panels of Figure 1 nonparametrically plot the relationship between expenditures and variety in India *within* village/block for 2009-2010, for food varieties and for grains and vegetables separately.

While this fact may seem obvious, there is no intrinsic reason that higher spending households must purchase more varieties. Richer households could consume the same number of varieties but purchase higher quantities or substitute cheap for expensive varieties. While this may occur, the data overwhelmingly support the notion that higher spending households spread their expenditure across a larger number of varieties. This mechanism seems especially important for food consumption in developing countries where poor households consume monotonous, staple-heavy diets and richer households diversify. However, the positive relationship between expenditures and variety within a location also holds for countries like Spain (within food and overall), the United Kingdom (food), and the United States (Nielsen Colorado sample, consumer packaged goods) using different survey methods, definitions of variety, and geographic aggregations (Appendix Figure A.1). Within the US as a whole, Broda and Romalis (2009) observe that households in higher percentiles of the expenditure distribution purchase more unique UPCs. This fact supports the existence of theoretical mechanisms that increase variety for richer households such as diminishing returns to quantity per variety.

B. Fact 2: Household variety differs systematically across locations

The Indian data reveal systematic differences in household variety across locations and time periods conditional on household expenditure. The bottom 3 panels of Figure

1 present examples of the general pattern, with variety Engel curves shifted up in later periods and in urban (vs. rural) locations. Appendix Figures A.2 and A.3 show that the size of this effect varies substantially across regions.

These patterns suggest that the retail differences could be important in shaping household food diversity. Table 2 examines which location characteristics correlate with household food variety. I group households by district, the most disaggregated location that can be linked across data sets and over time. I pool data from the 43rd, 61st and 66th NSS rounds to define 289 consistent districts that I match to the ICRISAT VDSA data set. All regressions include time fixed effects and population density. Column 1 shows that household food variety has an elasticity with respect to household food expenditure of about 0.3 and smaller but significant elasticity with respect to household size. Column 2 adds various district-level controls and shows that household food variety is higher in richer districts and in those with higher road density and a higher share of the population employed in food retail, wholesale, and service. I also consider two variables capturing dispersion (coefficient of variation) of prices and within-household expenditure shares across varieties for a district. Both variables are negatively correlated with household food variety, suggesting that household variety is higher when varieties are more symmetric. Column 3 adds the total number of food varieties observed in a household's district and village/urban block, which capture the "market-level" variety usually measured in the empirical literature. Even conditional on these measures, there are systematic differences in household food variety correlated with individual household (expenditure) and location (retail environment) characteristics. Columns 4 and 5 include district fixed effects. While the magnitude and statistical significance of some variables are reduced, road density and food retail remain significant predictors of household food variety.

Table 2 confirms the intuition that more developed retail environments are conducive to greater food variety for households. The retail environment is an equilibrium outcome of aggregate demand and supply factors, including aggregate expenditures, but it is a conceptually distinct factor affecting variety consumption than the within-location variety Engel curve that captures the effect of individual expenditure holding retail constant. Variety differences between any two households can be separated into a component related to own expenditure (Fact 1, movement along the variety Engel curve) and a component related to the local retail environment (Fact 2, the shift in a variety Engel curve representing differences in variety holding own expenditure constant). The same analysis also applies when comparing differences in mean variety as in Table 1 although differences in mean expenditure can affect variety through both channels in general equilibrium.

The importance of the retail environment for household variety in the Indian context supports a modeling approach that flexibly captures location-based determinants of variety. Appendix Figure A.4 shows that shifts in the food variety Engel curve over time for Spain and across Colorado MSAs are much smaller than what is observed in India. This could be due to aggregation although Spanish varieties are more aggregated and US varieties less so than in the Indian data. Broda and Romalis (2009) find that despite growth in aggregate US variety, UPCs per household actually fell between 1998 and 2005 within each expenditure quintile (almost 12% for the poorest quintile). Thus household variety may not always increase over time or correlate positively with market level variety. Modern supermarkets, which facilitate access to many food varieties, may weaken the link between the retail environment and household food variety in developed countries.

C. Fact 3: Indian households move up a hierarchy of food varieties

The types of varieties consumed by poor and rich (low and high variety) households may be the same, partly overlap, or be completely different. To explore how the type of varieties consumed varies across households, for each region I rank the varieties along one of four dimensions: aggregate regional expenditure share, the fraction of households that consume the variety in the region (the extensive margin), the average regional expenditure conditional on purchase (the intensive margin), and the median regional unit value (interpreted as the price). For each household I then compute a variety composition index for each dimension by averaging the rank of the varieties consumed by the household. Ranking varieties at the region level requires averaging over both rural and urban locations in a region but the ranking of varieties is quite stable across locations and over time within a region. Appendix Figure A.5 plots the share of consuming households for each region and grain/vegetable variety and shows that the increase in variety over time and in urban locations is driven by a broad-based increase in the likelihood of consuming most varieties that largely preserves the initial ordering (rank correlations above 0.8).

Table 3 shows that the variety composition index for grains and vegetables varies systematically across households. For each dimension I regress the index on household variety, on expenditure and an urban dummy, and on all three. I rank varieties separately for each region and include a region dummy.[3] High variety households are more likely to consume varieties that are consumed by fewer households and in smaller quantities, but are more likely to consume varieties that are expensive (in rupees/KG) for grains, though not for vegetables. Varieties have different taste and nutrition characteristics, so prices may be less informative than expenditures for ranking their relative importance. Expenditure and urban location predict a more marginal variety composition, but conditional on the number of varieties consumed have a limited independent effect.

The goodness of fit in Table 3 does not support a purely deterministic hierarchy of varieties, but confirms that a first-order feature Indian grain and vegetable consumption is that high variety households add more marginal varieties to their consumption basket. This fact is context specific, as there are settings that feature inferior goods or one-for-one substitution of low-end for high-end varieties. Hierarchical consumption is unlikely when looking within a narrowly defined category where households typically purchase a single variety (e.g. toothpaste), or in settings where households exhaust their capability to consume more quantity or variety leaving quality substitution as the only margin. However, Faber and Fally (2017) show that households in the US generally agree on their

[3] Appendix Table A.2 shows similar results looking within village/block.

ranking of brands (in terms of relative expenditure shares) suggesting that a common ranking of varieties across households may be a more general feature of the data.

D. Fact 4: Household variety correlates with shopping effort

Consuming more varieties may entail costs beyond the opportunity cost of spending more per variety. These include non-monetary and monetary costs of acquisition and preparation and indivisibilities (e.g. minimum quantity requirements that are costly to overcome or equivalently bulk discounts). Some of these costs do not differ in obvious ways across retail environments. I focus on shopping effort here as it is easy to measure, quantitatively important, and robustly correlated with variety, expenditure, and the retail environment. Shopping patterns provide an additional motivation for the model: while rich and poor households in a location have the same choice set, the rich may expand the set of varieties they consume by exerting more effort, e.g. by visiting more retail outlets or distant ones, complicating inference on choice sets from household data.

India's Time-Use survey measures time spent "shopping for goods and non-personal services" and "travel related to household maintenance, management and shopping." This includes shopping for non-food but the prominence of food in the household budget (over 60% of expenditure on average) and high purchase frequency suggest that most shopping is food-related. The survey covers all individuals age 6 or older for 18,589 households in six Indian states in 1998-1999. It records time-use in 20 minute intervals over the previous 24 hours and up to two variant days (e.g. weekends/market days) which I use to estimate average daily time-use.

Table 4 examines the correlation of shopping time with household expenditures and the retail environment. Column 1 regresses daily shopping time on expenditure, household controls (listed in the table notes) and village/block fixed effects. Within a narrow location, richer households spend more time shopping. Column 2 shows that households that live in a big town (population over 200,000) or smaller urban area spend more time shopping than those in rural areas, conditional on expenditure. Column 3 adds mean expenditure and share of population employed in food retail by sector for the 26 sample districts (I also control for population density), and shows that households shop more in area with more food retail conditional on expenditure. Households shop more in areas with more food retail. The survey does not measure food variety or food expenditure, but linking it to 1999-2000 NSS data at the district/sector level, Column 4 shows that mean food variety, conditional on own expenditure and mean district food expenditure, strongly predicts shopping time. Households with easier access to variety shop more, not less, which is counterintuitive but reflects that the total variety cost incurred could be elastic with respect to the cost per variety.[4]

In Appendix Table A.3 I present supplemental results from the survey. The shopping time pattern is robust to using the extensive margin, restricting to time-use by the head of household's spouse, and accounting for live-in servants. Results for other time-use categories rule out some alternative interpretations and mechanisms. Travel for non-shopping reasons displays no correlation with expenditure or retail environment/variety. Cooking, a potential variety cost, rises with expenditure and food retail (with a much smaller elasticity than shopping) but not with local variety. Leisure rises with income (with a much smaller elasticity than shopping time) but is not correlated with local retail or variety. Thus even though richer households have a higher opportunity cost of time, the marginal benefit of time spent shopping (or cooking) to increase variety rises more rapidly with expenditure than the marginal benefit of leisure. Grain processing and free collection are negatively correlated with expenditure and food variety. Appendix Table A.1 shows that the results in Table 2 are unaffected by accounting for part-time servants, which behave like a monetary cost of variety and can be interpreted as such if they save household time by doing the shopping.

Shopping time differences potentially provide a lower bound on the benefits of household variety in a revealed preference framework. If we assume that (a) households can

[4] Couture (2015) finds a similar pattern for US restaurant density and travel times.

consume the same number of varieties for the same shopping time and (b) shopping time is the only cost of variety and only serves to increase variety, then the foregone income from shopping time provides a lower bound on the welfare gains from variety. Sample households work about 90 hours a week, so the one extra hour on average for urban (vs. rural) or 90th percentile retail density districts (vs. 10th) implies a variety benefit of at least 1.1% (1/90) of expenditure.

In the Appendix I also show that similar patterns occur in the US. Nielsen Homescan data (Appendix Table A.4) allow me to show directly that households that consume more UPCs undertake more shopping trips, primarily because they visit more unique stores. The American Time-Use survey (Appendix Table A.5) shows that household income and urban location both strongly predict higher shopping time overall and for groceries only. Previous research on the US shows that greater shopping effort by households with a lower opportunity cost of time, due to retirement or unemployment, can reduce the price paid for the same good (Aguiar and Hurst (2007), Kaplan and Menzio (2015)) but these studies hold the consumption basket fixed. My findings provide support for the idea that shopping time is a critical input into the number of varieties in the consumption basket and is hence related to total expenditure and the retail environment.

2. Theory

This section presents a model of variety demand with three ingredients – diminishing returns to quantity, a cost of variety that can vary across locations, and asymmetric/hierarchical varieties – that can be used to decompose the source and cost-of-living implications of differences in household variety. After describing the model I contrast it with two alternative classes of models with similar features.

A. Variety Engel curve model

A consumer with constant elasticity of substitution preferences maximizes

$$\max_{q_i} \left(\sum_{i \in \Omega} z_i^{\frac{1}{\sigma}} q_i^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \quad s.t. \sum_{i \in \Omega} p_i q_i \leq X \tag{1}$$

where σ is the elasticity of substitution, p_i and q_i are price and quantity, z_i is a demand-shifter like taste or quality, X is total expenditure, and the choice set of varieties is Ω. The demand function for variety j is $q_j = \frac{X}{p_j}\left(\frac{p_j}{P(\{p_i|i\in\Omega\})}\right)^{1-\sigma} z_j$, where $P(\{p_i|i \in \Omega\}) \equiv (\sum_{i\in\Omega} z_i p_i^{1-\sigma})^{\frac{1}{1-\sigma}}$ is the CES price index. The expenditure (cost-of-living) function associated with utility level U_0 is $X(U_0, \{p_i|i \in \Omega\}) = U_0 P(\{p_i|i \in \Omega\})$. When $\sigma > 1$, there is diminishing marginal utility of quantity per variety. By spreading expenditure across more varieties, consumers increase utility, and a consumer purchases all varieties in Ω.

Given that rich households consume more varieties than poor households and consumers rarely purchase all of the varieties observed in a location, individual households cannot be interpreted as CES consumers. However, incorporating a cost per variety can help match the data and capture a continuous notion of variety access. If the cost per variety is the same for households in a location, richer households, who consume higher quantities, will incur greater costs to diversify their consumption. Expansion of the village store could lower local variety costs and incentivize poor households to add a new variety to their basket, while benefiting richer households (who previously bought that variety at a distant market town) through greater convenience.

I formalize this intuition by assuming that each location (l) features a variety cost given by $VC_l(n) = F_l n^{\epsilon_l}$ where n is the number of varieties consumed and F_l and ϵ_l are location specific parameters of the variety cost function. If $\epsilon > 0$ then consuming more variety is costly, but the marginal cost of variety can be decreasing, constant, or increasing in the number varieties consumed. In principle variety costs can be modeled more flexibly with arbitrary costs for each variety (with arbitrarily high costs capturing "unavailability") but this simple parameterization is attractive for my application using variety Engel curves. Variety costs can be modeled as utility cost or budget costs. I use a utility cost formulation below because of the evidence on shopping presented earlier, but in the Appendix I show that this is isomorphic to budget costs under certain assumptions. The precise source of variety costs is not important for my application of the model provided that variety cost parameters are interpreted as location characteristics.

Because the data are consistent with a consumption hierarchy, I make an additional assumption to parameterize the extent to which the marginal varieties consumed by high variety households are less valuable than infra-marginal varieties. Specifically, I assume that varieties can be indexed from highest to lowest benefit on a continuum from $[0, n]$, with $z_i p_i^{1-\sigma} \equiv [1+(1-\sigma)/\theta](bi^{1/\theta})^{1-\sigma}$. Recall that p_i is the price of a variety and z_i captures anything else (taste, quality, etc.) that shifts demand up or down given price. This says that varieties should be ranked in terms of their relative expenditure shares rather than prices. Relative expenditure shares for varieties i and j, given by $z_i p_i^{1-\sigma}/z_j p_j^{1-\sigma}$ (which I hereafter call the "relative intensive margin"), are inversely proportional to their relative rank $(i/j)^{\frac{1-\sigma}{\theta}}$. The parameter θ captures asymmetry across varieties (i.e. the extent to which marginal varieties are less valuable due to price, quality or taste) and as $\theta \rightarrow \infty$ varieties become symmetric and have identical expenditure shares.

By adopting this parameterization, the CES price aggregator can be written as:

$$P_l(n) = b_l n^{-\psi_l},\ \psi_l \equiv \frac{1}{\sigma-1} - \frac{1}{\theta_l} \tag{2}$$

where n is the number of varieties consumed. The parameter ψ measures the (negative) elasticity a CES price index with respect to n and includes both the elasticity of substitution σ (which I will assume is constant across locations) and the asymmetry across varieties θ_l (which may vary across locations). The restriction $\theta > \sigma - 1$ is required for $\psi > 0$.[5] The parameter b_l shifts the price level of all varieties across locations without changing

[5] This parameterization is borrowed from Arkolakis et al. (2007) who use the parameter ψ to denote the curvature of the CES price index with respect to variety. In my application, it is the combined price and taste term that is assumed to be distributed Pareto across varieties. It is not necessary for p_i and z_i to share the same ranking or any particular correlation across varieties, as can be seen in Table 3 comparing grains versus vegetables.

the relative value of marginal varieties and also captures differences in the price level for households that consume a single common variety.

With this parameterization of variety costs and benefits, the household problem can be solved recursively in two stages. In the second stage, the household takes the number of varieties purchased as given and solves a standard CES problem by allocating expenditures across a given set of varieties (equation 1). In the first stage, the household chooses the optimal number of varieties to purchase by comparing the marginal benefits and costs. The first stage variety choice problem is:

$$\max_{n} U = \frac{X}{b_l n^{-\psi_l}} - F_l n_l^{\epsilon} \tag{3}$$

which has a unique interior solution (when $\epsilon_l > \psi_l > 0$) given by:

$$n^* = \left(\left[\frac{X}{b_l} \right] \frac{\psi_l}{F_l \epsilon_l} \right)^{\frac{1}{\epsilon_l - \psi_l}} . \tag{4}$$

The model is summarized by only six parameters – household expenditure (X), the CES elasticity parameter (σ), and four location parameters reflecting the relative benefits and costs of marginal varieties ($b_l, \theta_l, F_l, \epsilon_l$) – but is flexible enough to capture a positive log-linear variety Engel curve within location and shifts across locations driven by differences in the marginal benefit or cost of variety.

To build intuition for the the model, consider three equations: the first-order condition equating the marginal benefit and cost of variety (in logs):

$$\underbrace{log(\psi) + (\psi - 1)log(n) + log(X/b)}_{log(MB)} = \underbrace{log(\epsilon) + log(F) + (\epsilon - 1)log(n)}_{log(MC)} \tag{5}$$

the log-linear variety Engel curve (equation 3 in logs):

$$\ln n = \left[\frac{1}{\epsilon - \psi} \ln(\frac{\psi}{bF\epsilon}) \right] + \frac{1}{\epsilon - \psi} \ln X \tag{6}$$

and the relative intensive margin, i.e. relative variety expenditures along the hierarchy.

Figure 2 presents comparative statics. First consider two households in the same location (and hence the same parameters b, ψ, F, ϵ). Household 2 has higher expenditures ($X_2 > X_1$) and due to diminishing returns to quantity has a higher marginal benefit of variety. The marginal benefit curve shifts up (Panel A), which results in movement along the location's variety Engel curve (panel B) and higher variety. The relative intensive margin is unaffected (panel C) because relative expenditures across varieties, conditional on consuming a variety, do not depend on total expenditures.

Next consider two households with identical expenditures ($X_1 = X_2$) in locations that differ because the marginal benefit of variety is greater in location 2 ($\psi_2 > \psi_1$). The increase in marginal benefit of variety in location 2 (panel D) implies that the household in location 2 consumes more varieties and incurs greater variety costs until marginal benefit and cost are equalized. The variety Engel curve in location 2 is shifted up for any level of expenditure (panel E). Higher ψ raises the intercept but also the slope of the MB and variety Engel curves, resulting in a larger (percentage) increase in variety for richer households. Relative expenditures on marginal goods increase with ψ (Panel F), which benefits the richer households that consume them. A decrease in b would result in a parallel shift for the variety Engel curve in panel E (or no shift if expenditures are divided by b on the X-axis) and no change in panel F.

Finally, consider two households with identical expenditures in locations that differ because marginal variety costs are lower in location 2 ($\epsilon_2 < \epsilon_1$), perhaps because more varieties are sold in the village shop relative to the district market. This shifts down and flattens the variety marginal cost function in location 2 (panel G). Given the same marginal benefit of variety but a lower marginal cost, the household in location 2 consumes more varieties. Lower ϵ shifts up the slope and intercept of the variety Engel curve (panel H), similar to the effect of higher ψ. The difference between these two comparative statics can be seen by comparing panels F and I. The reduction in variety costs has no effect on the relative expenditures for marginal goods in panel I. This distinction is critical for separately identifying ψ and ϵ. Note that a decrease in the F parameter would result in a parallel shift up in the variety Engel curve (panel H).

The model allows the difference in variety consumed between any two households (indexed by 0 and 1) to be decomposed as follows (see the Appendix for derivation):

$$\begin{aligned}\ln(\frac{n_1}{n_0}) = \underbrace{\left\{\frac{1}{\epsilon_1-\psi_1}\ln(\frac{X_1/b_1}{X_0/b_0})\right\}}_{expenditure\ difference} + \underbrace{\left\{\frac{1}{\epsilon_1-\psi_1}\ln(\frac{\psi_1}{\psi_0}) - \ln(\frac{X_0\psi_0}{\epsilon_0F_0b_0})[\frac{1}{\epsilon_1-\psi_1}-\frac{1}{\epsilon_1-\psi_0}]\right\}}_{relative\ intensive\ margin\ difference} \\ + \underbrace{\left\{\frac{1}{\epsilon_1-\psi_1}\ln(\frac{\epsilon_0F_0}{\epsilon_1F_1}) - \ln(\frac{X_0\psi_0}{\epsilon_0F_0b_0})[\frac{1}{\epsilon_1-\psi_0}-\frac{1}{\epsilon_0-\psi_0}]\right\}}_{variety\ cost\ difference} \quad (7)\end{aligned}$$

where n_0 and n_1 are the number of varieties consumed by two households. The three bracketed terms in the decomposition capture differences in expenditure (holding constant variety marginal cost and the relative intensive margin), the relative intensive margins (holding constant variety costs and expenditure), and the variety cost function (holding constant variety benefit). Variety differs within-location only due to expenditures, but across locations the retail environment parameters (b, ψ, F, ϵ) also matter.

The welfare implications are summarized by the expenditure (cost-of-living) function:

$$X(U_0, \eta, b, F) \equiv U_0^{1-\eta} b F^{\eta} \Psi, \;\; \eta \equiv \frac{\psi}{\epsilon} \; and \; \Psi \equiv \left[\eta^{\eta/(1-\eta)} - \eta^{1/(1-\eta)}\right]^{\eta-1} \quad (8)$$

The expenditure function is homogeneous of degree one in prices through b. The expenditure required to achieve utility U_0 is lower when the average price level is lower, when marginal varieties are more valuable, and when the cost of variety is lower (lower F or ϵ).

While differences in b or F lead to proportionate changes in the cost-of-living for rich and poor, decreases in ϵ and increases in ψ, which lead to steeper variety Engel curves, disproportionately benefit the rich. As variety offsets diminishing returns to quantity, a steeper variety Engel curve implies more welfare inequality for a given level of expenditure inequality. Comparisons of the cost-of-living across locations depend on the reference utility and some locations may be more favorable to the rich or poor.

The variety cost incurred by a household is given by:

$$F(n^*) = \left(\frac{X}{b}\frac{\psi}{\epsilon F}\right)^{\frac{1}{\epsilon-\psi}} F \tag{9}$$

where n^* is the optimal choice of n given X and location parameters b, F, ψ, ϵ. If variety costs are observable, the model makes two testable predictions about these costs. First, *within a location* households with higher expenditures incur greater variety costs. This is exactly the pattern documented earlier for shopping times. Second, when comparing households with similar expenditures, those in locations with higher variety may incur greater variety costs. This is trivially true for the ψ location parameter but is also the case for the F and ϵ location parameters when the ϵ parameter is high (implying an elasticity of $F(n^*)$ with respect to these parameters above one). Even if the cost per variety is lower in a location, when households consume more varieties in response their total variety costs could be higher. This is exactly the pattern documented earlier for shopping times in India and the United States, where households in high variety retail environments spend more time shopping conditional on expenditure.

B. Alternative models

Two other classes of models have been used to analyze the purchase of multiple varieties: multiple discrete choice (Hendel (1999), Dube (2004)) and bounded marginal utility (Wales and Woodland (1983), Kim et al. (2002)). Here I provide a brief overview of how these models compare to mine and how they fit the facts about household variety, leaving a fuller exposition and the technical details for the appendix.

Multiple discrete choice models assume that households choose a single variety from a choice set on a given purchase/consumption occasion but there are multiple occasions in the data. If household tastes vary randomly across occasions and there are diminishing returns to quantity per occasion, we may observe households that purchase multiple varieties. Both the number of occasions and the discrete choice problem affect household variety. Holding constant the number of occasions, the potential for repeating varieties across occasions allows the model to generate a correlation between household variety and aggregate variety or expenditure share symmetry, as in Table 2. Utility is increasing in the number of occasions because households can counteract diminishing returns per occasion and because they get more taste draws. Households consume on all occasions unless there is a minimum quantity requirement, so rich and poor households in the same location consume the same number of varieties on average and do not exhibit hierarchical consumption (Table 3) unless the discrete choice is non-homothetic or rich households have more occasions. With an "occasion cost" richer households would optimally choose more occasions making the model closer to mine, but in existing implementations the number of occasions is typically not modeled as a choice variable and is independent from the discrete choice problem by assumption. Consequently, welfare gains from variety in the model result from the interaction of heterogeneous tastes with the market-level choice set and household variety itself has no additional welfare implications.

Quadratic utility and translated additive preferences have diminishing returns to quantity across varieties so household variety does have distinct welfare implications in these models. They could incorporate variety costs but (unlike CES) do not need variety costs to generate an expenditure varying reservation ("choke") price for each variety due to bounded marginal utility as quantity approaches zero. In the Appendix I show how simple versions of these models with a similar parameterization of prices to the variety cost model also generate log-linear variety Engel curves. Differences in relative prices or tastes for varieties across locations can generate shifts in the variety Engel curve and relative intensive margin but there is no parameter that shifts them independently. The same first-order condition for quantity pins down whether a variety is consumed and how much is consumed. Simple versions of these models thus resemble my model with variety cost parameters held fixed across locations, with shifts in variety Engel curves tied to shifts in relative quantities. This is limiting in practice when the data feature opposite movements in the extensive and intensive margins at the level of individual varieties. For example, between 1983 and 2009 about 25% of varieties were consumed by a larger share of households despite lower aggregate quantity at the region level (Appendix Figure A.6). My estimation results imply that shifts in the relative intensive margin are not that predictive of variety and in some cases go in the wrong direction.

More elaborate versions of these models – with tastes or occasions that vary with expenditure or with varieties defined at a granular level (e.g. retail location) – can more flexibly fit the facts presented earlier. The key difference with my approach is not the presence of variety costs per se but their role in estimation. My approach imposes parametric structure on the benefits of variety to recover variety cost parameters capturing the (unobserved) location-determinants of household variety from household data. The other models have been estimated in contexts where one first observes the choice set and then estimates the parameters governing household variety. With detailed data on retail locations, assortments and shopping trips, a modeling approach that leverages this information has clear advantages over an indirect approach. One could explore counterfactuals like how the arrival or disappearance of a particular variety or retail outlet affects household variety or assess the relative importance of heterogeneous tastes or diminishing returns to quantity for welfare gains (e.g. Kim et al. (2002)). Data requirements make this approach infeasible in most settings. Market-level data on varieties can obscure the interaction of retail convenience and household expenditure which shape the endogenous choice of retail locations visited. Table 2 shows that this is the case even when measuring market-level variety at a very disaggregated level where household and "market" almost overlap, and in the next section I show systematic differences in the welfare gains from variety implied by my model versus a choice set approach.[6] Thus the primary advan-

[6] Handbury and Weinstein (2015) use accumulation curves to estimate market-level

tage of my model is its applicability to widely available household data without the need for prior identification of choice sets. It reduces differences in the retail environment to two variety cost parameters (F, ϵ) and one variety benefit parameter (ψ) estimated from household data on variety choice. My estimated variety cost parameters are correlated with measures of retail density but retail data is not required to estimate them.

4. Estimation and welfare

To estimate the variety cost model using Indian NSS data, I first estimate σ – the elasticity of substitution – using variation in the relative price and expenditure share of varieties over time. I then estimate ψ (equivalently θ) – the relative importance of marginal varieties – using variation in the relative expenditure share of marginal varieties across households. Finally, I recover location variety cost parameters, F and ϵ, from the slope and intercept of variety Engel curves given the other parameters.

My application considers two food groups – grains (18 varieties) and vegetables (29 varieties) – that make up a large share of the budget and feature varieties that look reasonably substitutable. Almost every household consumes at least one variety from these groups, they exhibit hierarchical consumption (Table 3), feature meaningful differences in variety across households and locations, and report quantities (critical for measuring prices and price elasticities). The cost-of-living differences I estimate are in terms of group expenditures only, although a first-order approximation of the overall welfare gains can be derived using the budget shares of these groups (28% for grains and 5% for vegetables in 1983 – Appendix Figure A,7 shows that the share is falling with income for grains but flat for vegetables). I discuss the potential endogeneity of group expenditure below.

Except for the σ parameter, estimated at the group level, the other parameters (θ, b, F, ϵ) are estimated separately for each group and location. For comparison over time (1983 vs.

choice sets from a sample of households but their approach abstracts from the willingness of households with larger benefits from variety to travel greater distances and differences in the convenience of accessing greater variety.

2009), I use NSS regions as locations. For comparisons between rural and urban (in 2009), I use region by sector as a location.[7] Regions contain enough sample households to estimate the parameters with some precision, unlike smaller geographic units like districts or villages/urban blocks (10 households) that capture finer grained differences in the retail environment but generate much noisier parameter estimates. The 75 NSS regions can be mapped consistently over time and consist of contiguous districts in a state that share similar geography, rural population densities and cropping patterns. Comparing within regions captures much of the within-India variation in taste and dietary patterns (see Atkin (2013)) while still revealing large differences in the model parameters.

A. Price elasticity of demand (σ)

To estimate σ, I use the fact that, conditional on consuming two varieties, relative household demand for variety i can be expressed:

$$\ln\left(Share_{hi}/Share_{h0}\right) = \alpha_{ri} + \underbrace{\beta}_{=1-\sigma} \ln\left(p_{rti}/p_{rt0}\right) + \underbrace{u_{hi}}_{\ln(z_{hi}/z_{h0})} \tag{10}$$

where i is a variety, h is a household, and t and r are the survey year and region. Both expenditure shares and prices are in log differences relative to a base variety (denoted 0) which I define as the most widely consumed variety in the region to maximize sample size. The constant α is allowed to vary for each r by i combination and captures time-invariant regional differences in the taste for each variety relative to the base variety. Identifying variation comes from differences in the relative price over time within a particular variety and region. The error term u_{hi} can be interpreted as idiosyncratic variation across households in the taste terms z_i and z_0 in equation 1.

[7] Locations that were rural can be re-classified as urban in later years due to changes in population, incorporation, etc. across decennial censuses (Hnatkovska and Lahiri (2016)). Re-classification would lead to understatement of variety growth over time within each sector, so I consider regions with fixed boundaries for comparisons over time.

Estimation of σ by OLS is subject to several biases. Prices are measured using median unit values (expenditure divided by quantity), leading to sampling and measurement error. Changes in prices over time may also be correlated with changes in tastes. Upward sloping supply curves imply that higher relative demand for a variety is associated with a higher relative price (biasing σ down), while a positive correlation between tastes and productivity in the long-run combined with trade frictions (e.g. a home market effect or the home bias in Atkin (2013)) imply the opposite (biasing σ upwards).

I address this with two instruments for relative prices that plausibly shift relative supply but not demand: local rainfall and prices in neighboring markets. I interact regional rainfall with a dummy variable for each variety, based on the idea that some varieties are more sensitive to local rainfall than others. The first-stage results (Appendix Table A.6 and Appendix Table A.7) show that this is indeed the case. The overall strength of the instrument set is fairly high for vegetables (just below 10) but low for grains (just below 6). Vegetables are typically grown locally while grain production is more concentrated in a few states and grain prices are heavily regulated by the government (partly to offset the effect of weather shocks). The price in neighboring markets is constructed as the average price in other regions of the same state. This is a valid instrument if the source of endogeneity is measurement error or if relative demand shocks in other regions are uncorrelated with demand shocks in the home region. The first-stage for this instrument is very strong for both groups, but the exclusion restriction is less plausible.

As equation 10 uses household expenditure shares, sample selection is another potential source of bias. Changes in market-level expenditure shares include both extensive (household) and intensive (within-household) changes in expenditure but the model elasticity is within-household. This is only a concern if the idiosyncratic component of variety tastes is realized before households choose their varieties, but if so, households with lower relative taste for a variety enter the sample when its price falls, resulting in a negative correlation between relative prices and u_{hi}. Lacking household panel data, I examine this issue in two ways. First, I control for the share of households consuming the variety. If the average taste for a variety falls as the share of households consuming it rises, this variable should absorb some of the taste variation with a negative coefficient. Second, I restrict the sample to varieties consumed by at least two-thirds of households, which excludes most varieties but leaves enough for estimation.[8]

Table 5 presents estimates using rural areas of the 75 NSS regions in 1983 and 2009-2010. Because the rainfall instruments are on the weaker side, I estimate use continuously-updating GMM which is efficient under clustering/heteroskedasticity but more robust to weak instruments. The IV elasticities are generally larger than the OLS elasticities, significantly so in terms of a Hausman test, but all specifications yield estimates of σ significantly above one. Accounting for sample composition/selection only has small effects on the IV point estimates (relative to standard errors), alleviating concerns that this is a quantitatively important source of bias. My preferred estimates control for the share of households consuming and use the rainfall IV for vegetables ($\sigma = 1.99$, $s.e. 0.098$) and neighboring regions IV for grains ($\sigma = 2.16$, $s.e. 0.128$) given instrument strength and plausibility. As the next steps in the estimation use σ to back out other model parameters region-by-region, I draw a σ parameter from a normal distribution with mean and standard deviation corresponding to these estimates when boot-strapping standard errors.

B. Other model parameters (θ, b, ϵ, F)

While I estimate a common σ, the parameter governing the relative importance of marginal varieties ($\psi \equiv \frac{1}{\sigma-1} - \frac{1}{\theta}$) may vary across locations due to θ. Expenditure on a

[8] A distinct issue is that elasticities may vary across rich and poor households. Appendix Table A.8 adds an interaction term for households with above median expenditures. The interaction term is often positive, suggesting that richer households have lower elasticities of substitution as in Handbury (2013), but these terms are small in magnitude and not significantly different than zero. Along with the results for widely consumed varieties, this provides support for the constant elasticity of substitution assumption.

base variety 0 can be written $x_{0h} = X_h \left(\frac{p_0}{bn_h^{-\psi}}\right)^{1-\sigma} z_{0h}$. Re-arranging and taking logs yields the estimating equation

$$\ln(X/x_0)_h = \underbrace{\alpha}_{[\ln(p_0/b)]} + \underbrace{\beta}_{[(\sigma-1)\psi]} \ln n_h + \underbrace{u_h}_{z_{0h}} \tag{11}$$

which I estimate separately for each location to derive ψ given an estimate of σ. This relationship reflects the proportionality of variety benefit to expenditure share in the model. If the expenditure share of a base variety falls more rapidly for high variety households, it implies a higher marginal benefit of variety and higher ψ. Under perfect symmetry, $\psi = \frac{1}{\sigma-1}$ and β above is one, e.g. increasing variety by 1% lowers the expenditure share of each variety by 1%. Appendix Figure A.8 plots an example of this approximately log-linear variation in the data. The estimation does not require a common ordering of varieties across households. The key assumption is that the valuation of the base variety is the same across households on average, which is required to measure the relative value of non-common/marginal varieties and is similar to the assumption in Feenstra (1994) that the taste for common varieties is the same.

With an estimate of ψ, I can back out the remaining location parameters using the log-linear variety Engel curve:

$$\ln n_h = \underbrace{\alpha}_{\left[\frac{1}{\epsilon-\psi}\ln(\frac{\psi}{bF\epsilon})\right]} + \underbrace{\beta}_{\frac{1}{\epsilon-\psi}} \ln X_h + u_h \tag{12}$$

The ϵ variety cost parameter can be recovered from the slope, while the F variety cost parameter can be recovered from the intercept given the other parameters and b (discussed below). The error term u_h can be interpreted as idiosyncratic variation in the F parameter within a location (e.g. $\ln F_h = \ln F + u_h$, $u_h \sim i.i.d.\ N(0, \sigma_u^2)$).

For both regressions, I include controls for household attributes that may affect variety through prices, tastes or variety cost: log household size, the share of male and female adults, dummies for agricultural workers, scheduled caste and scheduled tribe, religion, Public Distribution System consumption (grains) and the share of consumption produced by the household. Given the limited number of households per region used for estimation, I do not estimate heterogeneity in ϵ or ψ across households, but in principle this could be done with panel data or subgroups. For equations 11 and 12, I cannot reject equality of the slope or intercept above/below the median per capita expenditure for most regions (Appendix Figure A.11) which also supports log-linearity.

Both equation 11 and 12 require choosing a base variety. The choice matters because the benefits and costs of variety are measured relative to the base variety. Equation 11 is also not defined for households that do not consume the base variety which affects sample selection. If preferences are idiosyncratic, the variety consumed by the most households may not have the highest expenditure share, affecting the interpretation of equation 11. The parameter b in equation 12 is also necessary to recover the F variety cost parameter. Because $P(n) = bn^{-\psi}$ and $P(1) = b$, the parameter b governs differences in the price level across locations when comparing households that consume only a single variety.

One way around these issues is to use the preferred variety of each household as the base; in other words, X/x_0 is defined as expenditures on all varieties relative to the preferred variety of each household. Households are only dropped from estimation if they report zero expenditures for the group (about 3% of sample households). Because the base variety varies across households in this case, I set the parameter b equal to a Tornqvist price index defined over all varieties using aggregate expenditure shares.

I also estimate the model using the most widely consumed variety in a region as the base variety with b equal to its price. Although only 8% of households do not consume their region's preferred variety, when I exclude households for which this is not the highest expenditure variety I drop 22% of households for grains and 48% for vegetables. Appendix Figure A.9 shows that except for a few outliers the slope parameter estimates (ψ and ϵ) are similar across specifications. There are some differences in F that reflect its dependence on the b parameter, but these are minimal when comparing F within region. Appendix Table A.9 reports results using the household-specific and common base variety assumptions and the main conclusions are very similar.

I estimate equations 11 and 12 by OLS, which measure the slopes in the data. This estimates the structural parameters of the model in the presence of unobserved heterogeneity under a specific timing assumption: households first choose X, then receive IID disturbances to F that result in different choices of n (taking ψ as given), and then either receive a taste shock for the base variety relative to others (z_{0h}) or have IID measurement error in the dependent variable X/x_0. Measurement error in the independent variables (n for equation 11 and X for equation 12) or endogeneity due to violation of the timing assumption (e.g. households choose X after observing F or choose n after z_{0h}) could lead to biased estimates of the structural parameters. I provide a simple test for whether this matters using education of the household head as an instrument for both equations, as it strongly predicts group expenditure and variety while plausibly satisfying the exclusion restriction. Appendix Figure A.10 presents the distribution (across regions) of p-values from the Hausman test for equations 11 and 12 for both groups. For almost every region the Hausman test fails to reject the null. Given that the OLS estimates are already quite noisy I report results using the more efficient OLS estimates.

C. Decomposition and welfare results for variety Engel curve model

I present the broad patterns here and report the full set of results for all 75 regions in Appendix Tables A.10 through A.14. I generate standard errors by boot-strapping: for each location I draw 1000 random samples with replacement and a value of σ (from the distribution of estimates discussed above), each time estimating ψ, then ϵ and F, and then the decomposition and cost-of-living terms. Appendix Figures A.12 and A.13 plot the entire distribution of point estimates and t-statistics for welfare gains across regions.

Table 6 presents results for comparisons within-region between 1983 and 2009. In addition to the mean across 75 regions, I also report means for the bottom and top quintile of regions ranked by population density in 1983. Panel A presents the decomposition of variety changes into expenditure, relative intensive margin (ψ) and variety cost (F, ϵ) based on equation 7. The decomposition can be undertaken for any two households. Here I consider a comparison of households with the median group expenditure in each location. The median household experienced a 50% increase in variety in the average region. Variety increases for vegetables were highest in less densely populated regions, but variety increases for grains were slightly lower in these regions. Regions with lower per capita expenditures in 1983 also saw larger increases in variety for both groups during this period, consistent with some convergence of the poorest/most remote markets.[9] Most of the growth in variety over this period is driven by variety cost parameters. Expenditure growth played a small or even negative role for grains as expenditure fell in most regions, but contributed positively to vegetable variety in most regions and accounted for a substantial part of variety growth in low density regions. For grains, a decline in the ψ parameter contributed negatively to variety growth. This reflects a general trend away from coarse cereals and towards rice and wheat, the two predominant grain "base" varieties, due in part to government intervention through the Public Distribution System. For vegetables, the impact of changes in ψ on variety growth is positive but smaller.

Panel B presents welfare gains based on equation 8, which depends on the reference utility level. A location with higher ψ or lower ϵ appeals to all households but particularly those with higher expenditure. I choose the reference utility level corresponding to the household with median expenditure in 1983. Welfare gains are expressed as the share of group expenditure that this household would give up to face the parameters (ψ, F, ϵ) of the comparison location. I also report results that only factor in differences in ϵ, F (holding ψ constant) to highlight the effects of changes in variety cost parameters only. Because the

[9] Grain and vegetable variety increases have a correlation of -0.09 and -0.41 (significant at the 1% level) with population density over this period and a correlation of -0.22 with per capita expenditure (significant at the 10% level), with similar correlations for welfare.

"rich-biasedness" of welfare gains is of independent interest, I also report the differential gains for households at the 90th vs. 10th percentile of expenditure. These reflect how much more rich households are willing to pay relative to poor households, as a fraction of their expenditures, to face the parameters of the comparison location.

The average region experiences large welfare gains from better access to vegetable variety (25% expenditure equivalent). Gains from grain variety access are about half as large due to lower average ψ, i.e. households benefit less from accessing marginal grain varieties. Changes in ψ have a limited impact on changes in variety but because they affect welfare from infra-marginal varieties they can have substantial welfare effects, lowering welfare gains for grains to about 2% but raising them for vegetables to 30%. More densely populated regions saw slightly larger gains from access to grain variety but much smaller gains from access to vegetable variety. Welfare gains were rich-biased for grains but poor-biased for vegetables with a similar degree of bias across regions.

Table 7 presents analogous results comparing rural and urban areas within regions in 2009-2010. I report means for the bottom and top quintile of regions ranked by the urban-rural gap in mean per capita total expenditure. Panel A reports that the median urban household consumes 30% more grain and 10% more vegetable varieties. Most of the urban-rural gap is due to lower variety cost parameters in urban areas. Expenditure differences account for a quarter of the urban-rural variety gap for vegetables. The ψ parameter tends to be lower in urban areas which contributes negatively to urban-rural difference in variety. Regions with a larger urban-rural income gap exhibit larger gaps in urban-rural variety. For vegetables a substantial part of this is due to expenditure differences, but the majority is still due to variety costs, suggesting that regions with relatively poor rural areas also feature relatively low rural access to variety.

In the Panel B welfare comparison, lower variety cost parameters contributed to modestly higher welfare in urban areas overall (3.5-4.5%), with larger welfare differences in the regions with the largest urban-rural income gap. When differences in ψ are included in the cost-of-living comparison, the advantages of living in urban areas are reduced – marginal varieties are more accessible in urban areas but they have smaller expenditure shares. Welfare gains in the urban-rural comparison are rich-biased for both groups. Given that the urban-rural variety and welfare gap is largest for regions with the greatest urban-rural income disparities, these results suggest that consumption amenities may contribute to urbanization and sorting based on income.

D. Interpretation of variety cost parameter estimates

The results show that most of the differences in variety over time and across sectors are driven by variety cost parameters. This implies that differences in expenditure and the benefit of marginal varieties are not enough to explain observed variety differences (or even go in the wrong direction) but is silent on what explains these parameters. Appendix Table A.14 explores the correlation of the variety cost parameter estimates from Table 7 with proxies for the retail environment used in Table 2. Total variety costs (Fn^{ϵ}) depend on the two variety cost parameters and on the choice of n (which depends on these but also other variables), so I include a measure of variety costs that holds n constant at the urban median and one that uses the actual n for the median household in each region-sector. When holding n fixed, variety costs are negatively correlated with share of the population in food retail, but when allowing n to vary this correlation is reversed. The pattern is similar for other variables like mean per capita expenditure and population. The results for variety cost that allow n to vary are similar to those for shopping time (Table 4) and correspond to the model's prediction that household variety costs incurred are elastic with respect to the variety cost parameters.

While variety cost parameters are identified based on differences in variety across the expenditure distribution (i.e. the shift in variety Engel curves), for any given household the welfare gains from differences in variety cost parameters include both "gains from variety" (in the sense that lower marginal variety costs increase variety chosen) and lower infra-marginal variety costs for varieties already consumed. The total magnitude of infra-marginal variety costs in consumption equivalent terms is low. The median household in the average region would give up 4% of grain expenditures and 17% of vegetable expenditure to consume the same variety but face zero variety costs. The share of the welfare gains due to changes in F and ϵ in Table 6 that would have resulted holding n constant ranges from 11% for grains to 36% for vegetables. The reason these numbers are quite low is that I estimate high values for ϵ (typically above 4) to match the variety Engel curve slopes, implying that marginal variety costs rise rapidly with variety. Increases in observed variety for poor households, which identify changes in infra-marginal variety costs for richer households, require only small reductions in variety costs for richer households in consumption equivalent terms.

E. Welfare comparison with aggregate CES

Aggregate CES preferences using market-level variety provide an interesting comparison given model similarities. Recall that these capture the normative implications of single and multiple-discrete choice models with multinomial logit tastes as well as an aggregate consumer with diminishing returns to quantity. I apply the cost-of-living formula from Feenstra (1994) to households aggregated at the region or village/block level, using the same estimates of σ to highlight the importance of assumptions about variety access.[10] There are thousands of bilateral village/block comparisons per region so I report the median which should also mitigate sampling issues.[11]

Panel C of Tables 6 and 7 present the CES welfare results. Geographic aggregation is critically important. At the region level the overlap in varieties is almost complete resulting in tiny cost-of-living differences. At the village/block level there is less overlap and larger welfare effects. Table 2 showed that differences in village/block variety are

[10] Under multiple discrete choice, household and aggregate σ are the same as both reflect the same distribution of tastes whose variance pins down σ.

[11] With only 10 households per village/block, there are too few observations to implement the methodology in Handbury and Weinstein (2015).

highly correlated with household variety, so it is not surprising that the welfare gains at this level of CES aggregation are closer to my model. Under multiple discrete choice, a larger choice set can lead to higher average variety, while lower variety costs in my model can lead to higher observed market-level variety.

Deviations between welfare gains under village/block CES and my model display a systematic pattern: the village CES model leads to higher gains when variety differences and expenditures are positively correlated (vegetables) and lower gains when they are negatively correlated (grains). This highlights an important difference between household and aggregate variety concepts. Because expenditure affects household variety, even without changes in the retail environment, variety differences in a sample of households will tend to correlate with differences in the expenditures of the households in the sample. My approach does not rule out that aggregate expenditures affect the retail environment (variety cost parameters) in general equilibrium (e.g. see Appendix Table A.14) but it does imply that individual household expenditure is an independent factor that must be accounted for when estimating the contribution of the retail environment to variety and welfare differences in household data. The comparison shows that a simple CES approach, if applied at a disaggregated level, can broadly capture the same forces underlying differences in variety cost parameters, particularly when the data exhibit three features that are mostly satisfied in my setting – differences in expenditure levels across locations are small, the slope of variety Engel curves is not too high (ϵ is high), and differences in household variety are mostly driven by differences in variety access. However, estimates of choice sets derived from household data may still benefit from accounting for the role of household expenditures when higher expenditure households purchase more varieties and exert more shopping effort.

5. Conclusion

In this paper, I document large differences in the variety of goods consumed across households and analyzed the source and welfare implications of these differences. Household variety depends on both individual demand factors like expenditure as well as features of the local retail environment that vary over time and across locations. I develop a simple model in which the benefits of marginal variety, due to diminishing marginal returns to quantity, are balanced against a marginal cost of variety that captures (unobserved) aspects of the local retail environment. The model generates the log-linear variety Engel observed in the data, which can be used to recover the variety cost parameters relevant for cost-of-living estimation. Estimating the model on Indian household data, I show that while differences in the marginal benefit of variety due to expenditure or relative prices/tastes of marginal varieties matter for household variety, most of the increases in household food variety over time and in urban areas are driven by lower variety cost parameters that imply large welfare gains. I believe these are the first estimates of consumer welfare gains from dietary diversity in developing countries, where the transition from monotonous, staple-heavy diets to more diverse consumption appears to be a ubiquitous feature of economic development and urbanization. My estimates highlight consumption diversity as a potential motivation for rural to urban migration and suggest that increasing the efficiency of the retail and distribution sector in developing countries could benefit households by reducing the cost of variety.

An appealing feature of my modeling approach is that it estimates the contribution of location factors to household variety and the cost-of-living using only household data, which is widely available for many countries, many types of goods, and many geographic aggregations. However, data combining household purchase and shopping behavior with detailed retailer data could allow for the estimation of richer interactions between the retail and household factors that shape household variety consumption, as well as a more precise characterization of the mechanisms that limit variety consumption and show up as variety costs in my model. There are existing data sources in developed countries that may allow this, although it is an open question whether differences in retail environments in this context – where a single supermarket or on-line retailer puts thousands of varieties a few aisles or clicks away – are important for household variety. In developed countries location factors may matter more for non-tradable varieties like restaurants and live entertainment than basic food items. A better understanding of the relationship between household variety and the retail environment could also shed further light on the relative importance of diminishing returns to quantity versus heterogeneous tastes as channels through which households benefit from variety.

References

Aguiar, Mark and Erik Hurst, "Lifecycle Prices and Production," *American Economic Review*, 2007, *97(5)*, 1533–1559.

Allender, William J., Timothy J. Richards, Sungho Park, and Stephen F. Hamilton, "Demand for Variety Under Costly Consumer Search: A Multiple-Discrete/Continuous Approach," *Kilts Booth Marketing series Paper 1-002*, 2013.

Almas, Ingvild, "International Income Inequality: Measuring PPP bias by estimating Engel curves for food," *American Economic Review*, 2012.

Anderson, Simon P., Andre de Palma, and Jacques-Francois Thisse, *Discrete Choice Theory of Product Differentiation*, MIT Press, 1992.

Arkolakis, Costas, Svetlana Demidova, Peter J. Klenow, and Andres Rodriguez-Clare, "Endogenous Variety and the Gains from Trade," *Working Paper*, 2007.

Atkin, David, "Trade, Tastes and Nutrition in India," *American Economic Review*, 2013, *103(5)*, 1629–1663.

_ , Benjamin Faber, and Marco Gonzalez-Navarro, "Retail Globalization and Household Welfare: Evidence from Mexico," *Journal of Political Economy*, 2016.

Banerjee, Abhijit and Esther Duflo, *Poor Economics*, Public Affairs, 2011.

Bils, Mark and Peter J. Klenow, “Quantifying Quality Growth,” *American Economic Review*, 2001a, *91*, 1006–1030.

Broda, Christian and David E. Weinstein, “Globalization and the Gains from Variety,” *Quarterly Journal of Economics*, 2006, *121(2)*, 541–585.

_ and _ , “Product Creation and Destruction: Evidence and Price Implications,” *American Economic Review*, 2010, *100(3)*, 691–723.

_ and John Romalis, “The Welfare Implications of Rising Price Dispersion,” *Working Paper*, 2009.

Bronnenberg, Bart J., “The Provision of Convenience and Variety by the Market,” *RAND Journal of Economics*, 2015, *46(3)*, 480–498.

Couture, Victor, “Valuing the Consumption Benefits of Urban Density,” *Working Paper*, 2015.

Deaton, Angus, “Quality, Quantity and Spatial Variation in Price,” *American Economic Review*, 1988, *78*, 418–430.

Dube, Jean-Pierre, “Multiple Discreteness and Product Differentiation: Demand for Carbonated Soft Drinks,” *Marketing Science*, 2004, *23(1)*, 66–81.

Faber, Benjamin and Thibault Fally, “Firm Heterogeneity in Consumption Baskets: Evidence from Home and Store Scanner Data,” *Working Paper*, 2017.

Feenstra, Robert C., “New Product Varieties and the Measurement of International Prices,” *American Economic Review*, 1994, *84(1)*, 157–177.

Hamilton, Bruce W., “Using Engel’s Law to Estimate CPI bias,” *American Economic Review*, 2001, *91(3)*, 619–630.

Handbury, Jessie, "Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living Across U.S. Cities," *Working Paper*, 2013.

_ and David E. Weinstein, "Goods Prices and Availability in Cities," *Review of Economic Studies*, 2015, *82(1)*, 258–296.

Hendel, Igal, "Estimating Multiple-Discrete Choice Models: An Application to Computerization Returns," *Review of Economic Studies*, 1999, pp. 423–446.

Hnatkovska, Victoria and Amartya Lahiri, "Urban Sprawl and Rural Development: Theory and Evidence from India," *Working Paper*, 2016.

Hsieh, Chang-Tai, Nicholas Li, Ralph Ossa, and Mu-Jeung Yang, "Accounting for the New Gains from Trade Liberalization," *Working Paper*, 2016.

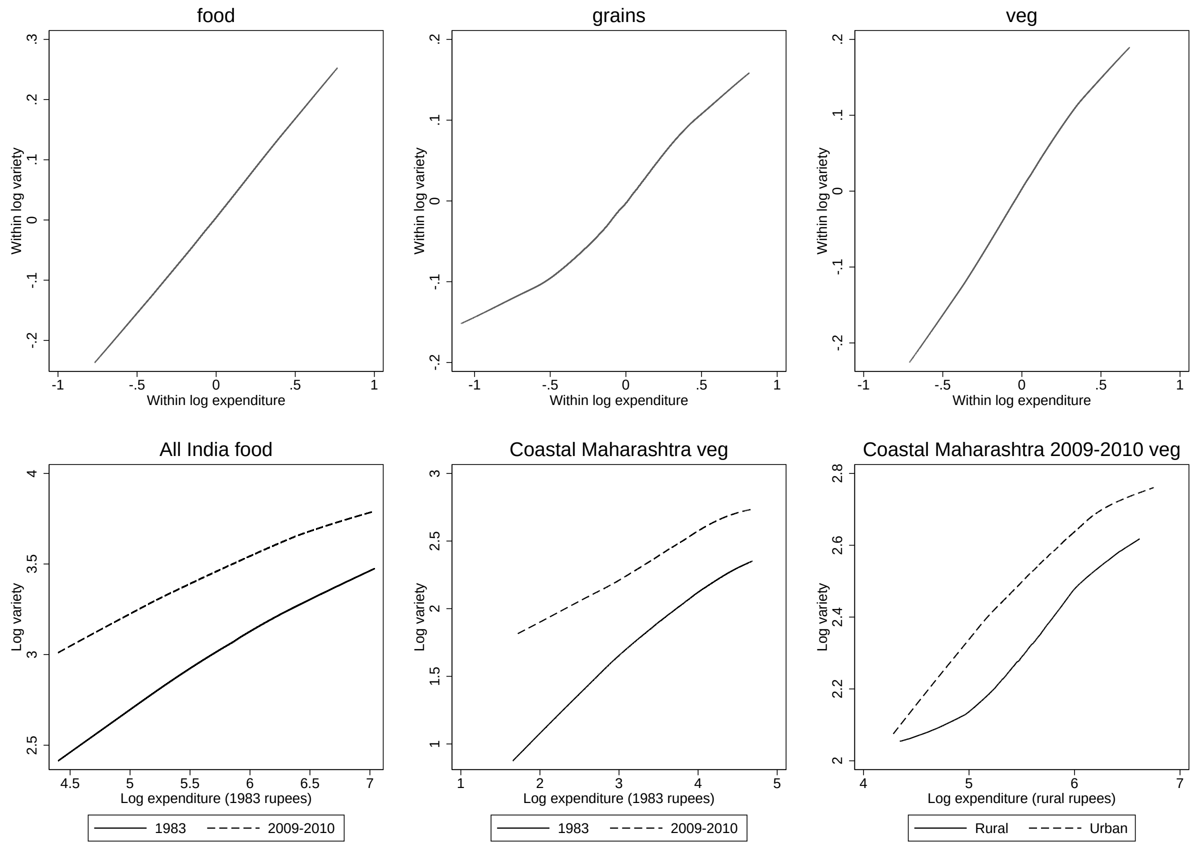
Kaplan, Greg and Guido Menzio, "The Morphology of Price Dispersion," *International Economic Review*, 2015, *56(4)*, 1–42.

Kim, Jaehwan, Greg M. Allenby, and Peter E. Rossi, "Modeling Consumer Demand for Variety," *Marketing Science*, 2002, *21(3)*, 229–250.

Lagakos, David, "Explaining Cross-Country Productivity Differences in Retail Trade," *Journal of Political Economy*, 2015.

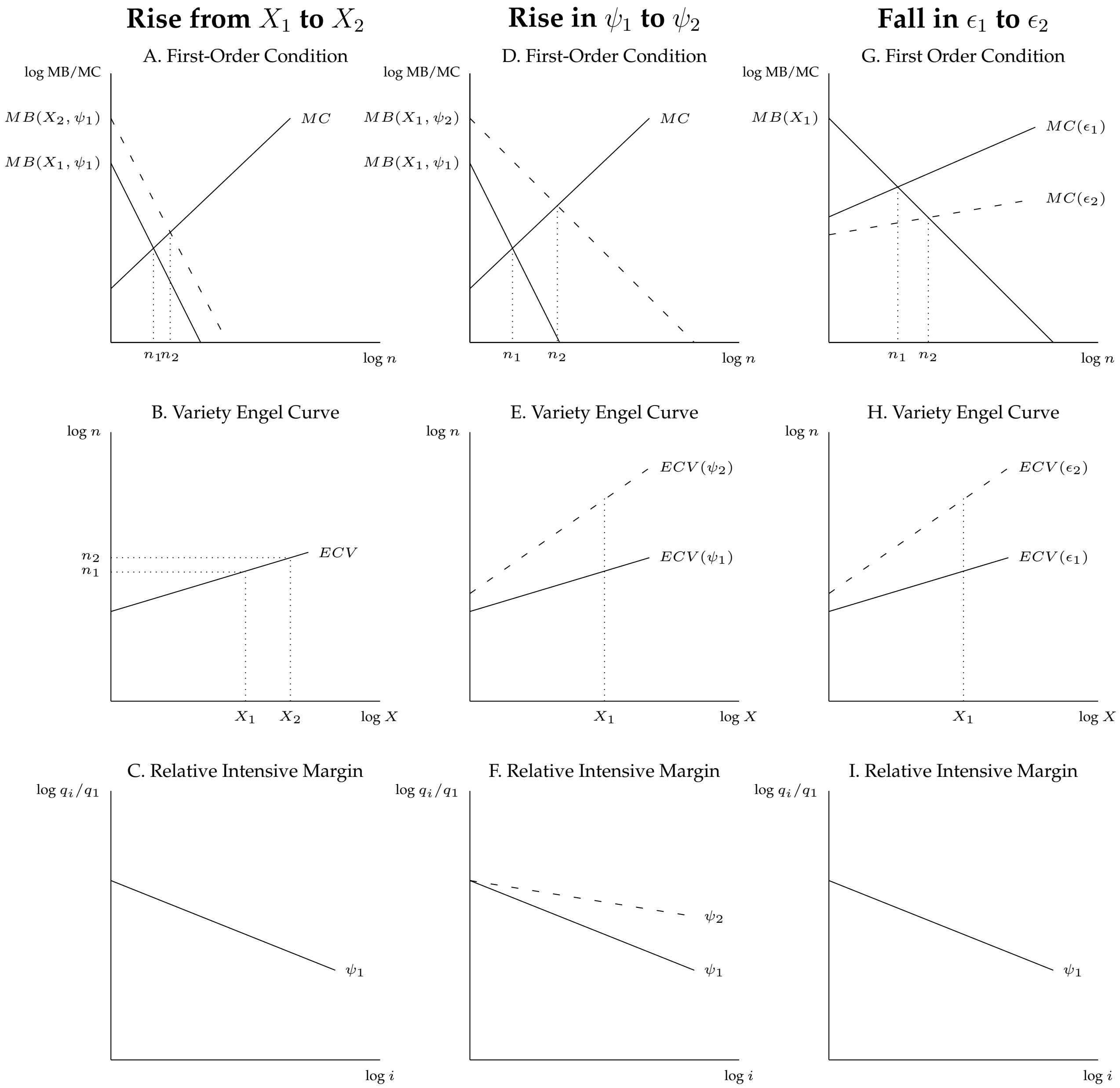
Wales, T.J. and A.D. Woodland, "Estimation of Consumer Demand Systems with Binding Non-Negativity Constraints," *Journal of Econometrics*, 1983, *21*, 263–285.

Figure 1: Fact 1 (variety increases in expenditure within location) and Fact 2 (variety differs across locations/periods for given real expenditures)



Note: Lowess regression of log household variety on log expenditure. Top 3 panels include all NSS households in 2009-2010 and partial out demographics, industry and village/block dummies. Bottom 3 panels use nominal expenditure deflated by a price index (with base equal to 1983 or rural areas).

Figure 2: Sources of variety growth: comparative statics



Rise from X_1 to X_2

Rise in ψ_1 to ψ_2

Fall in ϵ_1 to ϵ_2

A. First-Order Condition

D. First-Order Condition

G. First Order Condition

B. Variety Engel Curve

E. Variety Engel Curve

H. Variety Engel Curve

C. Relative Intensive Margin

F. Relative Intensive Margin

I. Relative Intensive Margin

Table 1: Variety and real expenditures by sector and year for India

Locations	All India		Rural	Urban
Year	1983	2009	2009	2009
Mean Variety Per Household				
Food (out of 134)	23.5	35.5	34.3	37.6
Grains (out of 18)	2.4	3.3	3.1	3.5
Vegetables (out of 29)	5.7	10.8	10.5	11.3
Mean Variety Per Region				
Food	127.3	129.6	124.7	126.7
Grains	16.6	15.9	15.2	14.8
Vegetables	28.2	28.6	28.1	28.3
Mean Real Expenditures				
Prices	1983 Rupees		2009 Rural Rupees	
Food	416	404	2522	2648
Grains	182	139	736	583
Vegetables	32	31	289	298
Households	117464	100855	59119	41736

Data from NSS rounds 38 and 66.

Grains: rice, chira, khoi/lawa, muri, other rice products, wheat/atta, maida, suji/rawa, sewai/noodles, bread(bakery), other wheat prod., jowar, bajra, maize, barley, small millets, ragi, cereal substitutes.

Vegetables: potato, onion, radish, carrot, turnip, beet, sweet potato, arum, pumpkin, gourd, bitter gourd, cucumber, parwal/patal, jhinga/torai, snake gourd, cauliflower, cabbage, brinjal, lady's finger, palak/other leafy, barbati, tomato, peas, chillis, capsicum, plantain, jackfruit(green), lemon, other.

Table 2: Household food variety, location characteristics and retail environment. Dependent variable is log number of varieties consumed by the household (Fact 2).

	(1)	(2)	(3)	(4)	(5)
Log real food exp.	0.339***	0.330***	0.265***	0.355***	0.291***
	(0.009)	(0.009)	(0.009)	(0.007)	(0.008)
Log household size	0.040**	0.062***	0.115***	0.058***	0.105***
	(0.015)	(0.017)	(0.017)	(0.018)	(0.018)
Log dist. mean exp.		0.119***	-0.023	0.001	-0.088***
		(0.037)	(0.021)	(0.030)	(0.025)
Log dist. road density		0.056***	0.027***	0.014*	0.006**
		(0.017)	(0.008)	(0.007)	(0.003)
Pop. share in food retail		5.583***	2.167***	1.180***	0.602*
		(0.649)	(0.352)	(0.395)	(0.346)
Log dist. price disp.		-0.057*	-0.044**	0.009	-0.018
		(0.030)	(0.020)	(0.020)	(0.015)
Log dist. exp. share disp.		-0.180***	-0.097***	-0.042	-0.060
		(0.049)	(0.029)	(0.069)	(0.052)
Log dist. variety			0.145***		0.037
			(0.043)		(0.063)
Log village variety			0.655***		0.548***
			(0.017)		(0.016)
District FE	No	No	No	Yes	Yes
R^2	0.359	0.400	0.493	0.473	0.521

N=195,586. Standard errors clustered by district. *** p<0.01, ** p<0.05, * p<0.1. Sample is households from 43rd, 61st, and 66th NSS rounds not missing any variables. Regressions include year fixed effects and population density.

Table 3: Variety composition depends on household variety (Fact 3). Dependent variable is average rank of varieties consumed by a household for a given regional characteristic (2009-2010 NSS).

Characteristic	Agg. exp. share			Frac. of hh (exp.>0)			Mean exp. if exp>0			Price		
Panel A: Grains (N=98,966)												
Num. varieties	0.68***		0.69***	0.63***		0.64***	1.34***		1.33***	0.79***		0.77***
	(0.02)		(0.02)	(0.01)		(0.01)	(0.05)		(0.05)	(0.04)		(0.04)
Log exp.		0.33***	-0.05***		0.30***	-0.05***		0.69***	-0.04*		0.41***	-0.02
		(0.02)	(0.01)		(0.02)	(0.01)		(0.04)	(0.02)		(0.04)	(0.03)
Urban		0.34***	-0.01		0.28***	-0.04**		0.80***	0.13***		0.75***	0.36***
		(0.03)	(0.02)		(0.02)	(0.02)		(0.07)	(0.04)		(0.06)	(0.04)
R^2	0.66	0.22	0.66	0.63	0.21	0.63	0.72	0.28	0.72	0.70	0.55	0.71
Panel B: Vegetables (N=98,813)												
Num. varieties	0.43***		0.44***	0.44***		0.44***	0.37***		0.41***	-0.00		-0.03
	(0.01)		(0.01)	(0.01)		(0.01)	(0.01)		(0.01)	(0.02)		(0.02)
Log exp.		1.09***	-0.17***		1.19***	-0.05*		0.76***	-0.39***		0.10	0.18**
		(0.04)	(0.03)		(0.04)	(0.03)		(0.05)	(0.05)		(0.06)	(0.07)
Urban		0.45***	0.12***		0.47***	0.14***		0.29***	-0.02		0.22***	0.24***
		(0.04)	(0.02)		(0.04)	(0.02)		(0.04)	(0.03)		(0.05)	(0.04)
R^2	0.63	0.28	0.64	0.64	0.30	0.65	0.54	0.34	0.55	0.44	0.45	0.45

Standard errors clustered by region. Price is median rupees/KG ranked low to high. See text for variable description.

Table 4: Household food variety and its predictors increase shopping time (Fact 4). Dependent variable is average household minutes per day shopping and traveling for shopping (mean=29).

	(1)	(2)	(3)	(4)
Log exp.	2.503*	1.925	3.059**	2.831**
	(1.429)	(1.581)	(1.400)	(1.302)
Other urban		7.528***		
		(2.012)		
Town pop>200K		13.44***		
		(3.748)		
Log mean exp.			7.869	
			(10.27)	
Pop. share in food retail			515.1**	
			(233.6)	
Mean food variety				1.178***
				(0.405)
Log mean food exp.				-2.120
				(5.524)
Village/block FE	Yes	No	No	No
R^2	0.642	0.049	0.050	0.057

N=18,589. Standard errors clustered by district. *** p<0.01, ** p<0.05, * p<0.1. Regressions include log household size, adult male and female ratios, dummies for caste, tribe, religion, farmer status, and house type. Time-use survey sample from 1998-1999 merged to NSS data for 1999-2000 for district/sector food variety and expenditure.

Table 5: Elasticity (σ) estimates: $\Delta \ln Share$ regressed on $\Delta \ln price$, reported coefficient is equivalent to $(1-\sigma)$

Specification	OLS			IV: rainfall x variety			IV: price in other regions of state		
Which varieties	All	All	Popular	All	All	Popular	All	All	Popular
Grain varieties (18 total)									
$\Delta \ln p$	-0.634***	-0.596***	-1.777***	-1.976***	-1.436***	-2.193***	-1.228***	-1.160***	-1.717***
	(0.0921)	(0.0805)	(0.223)	(0.231)	(0.202)	(0.0280)	(0.141)	(0.128)	(0.308)
Δ household share		1.467***			1.279***			1.418***	
		(0.199)			(0.198)			(0.203)	
Region-year-varieties	1370	1370	83	1142	1142	48	1142	1142	48
Vegetable varieties (29 total)									
$\Delta \ln p$	-0.108***	-0.0620**	-0.135	-1.028***	-0.988***	-1.522***	-0.448***	-0.363***	-0.600***
	(0.0312)	(0.0303)	(0.0893)	(0.0882)	(0.0984)	(0.415)	(0.0630)	(0.0557)	(0.180)
Δ household share		-0.468***			-0.194**			-0.383***	
		(0.0593)			(0.0838)			(0.0637)	
Region-year-varieties	2597	2597	357	2314	2314	136	2314	2314	136

*** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered by region-year-variety. Estimates are from equation 10 with region-variety fixed effects, using all rural households consuming at least two varieties in the 38th and 66th NSS rounds. IV estimates use continuously updating GMM. See Appendix tables A.6 and A.7 for first-stage coefficients and F-statistics.

Table 6: Region-level variety differences over time (mean across regions): 2009 vs. 1983 (base)

Group	Grains			Vegetables		
Region Pop. Density	All	Low	High	All	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Decomposition of difference in variety (2010 vs. 1983) at median						
Change in variety ($\%\Delta n$)	0.493	0.443	0.539	0.567	0.699	0.480
Expenditure component (X/b)	-0.046	-0.021	-0.031	0.047	0.079	0.021
Int. margin component (ψ)	-0.120	-0.181	-0.118	0.051	0.073	0.126
Variety cost component (F, ϵ)	0.661	0.627	0.688	0.469	0.547	0.333
Panel B: Welfare gains (2010 vs. 1983, as share of group expenditure) at median						
Change in F, ϵ only	0.135	0.128	0.138	0.253	0.269	0.204
Change in F, ϵ, ψ	0.017	-0.047	0.038	0.305	0.323	0.370
Gain at 90th vs. 10th pct. (F, ϵ)	0.019	0.020	0.010	-0.062	-0.051	-0.052
Panel C: Welfare gains (2010 vs. 1983, as share of group expenditure), alternative models						
Region CES	-0.009	0.000	-0.011	0.013	0.028	0.015
Village CES (median)	-0.009	0.007	0.000	0.333	0.429	0.216

Low and high density regions are the means for the bottom quintile (15) and top quintile (15) regions ranked by pop. density in 1983. Panel A is based on equation 7 and panel B is based on equation 8, evaluated at utility corresponding to median group expenditure in 1983. Rich bias is the difference between gains evaluated at 90th versus 10th percentile of group expenditure in 1983. Panel C is based on the Feenstra (1994) applied at the region or village/block level (median across all bilateral village/block comparisons over time within region).

Table 7: Within-region urban-rural variety gaps (mean across regions): urban 2009 vs. rural 2009 (base)

Group	Grains			Vegetables		
Region urban-rural inc. gap	All	Low	High	All	Low	High
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Decomposition of difference in variety (urban vs. rural) at median						
Change in variety ($\%\Delta n$)	0.292	0.140	0.436	0.093	0.023	0.204
Expenditure component (X/b)	-0.012	-0.020	-0.009	0.022	0.004	0.039
Int. margin component (ψ)	-0.026	-0.040	-0.012	-0.020	-0.018	0.011
Variety cost component (F,ϵ)	0.334	0.206	0.469	0.091	0.036	0.154
Panel B: Welfare gains (urban vs. rural, as share of group expenditure) at median						
Change in F,ϵ only	0.044	0.034	0.058	0.035	0.007	0.066
Change in F,ϵ,ψ	0.033	0.003	0.047	-0.049	0.004	-0.037
Gain at 90th vs. 10th pct. (F,ϵ)	0.025	0.022	0.040	0.017	0.010	0.024
Panel C: CES Welfare gains (urban vs. rural, as share of group expenditure)						
Region CES	-0.001	-0.001	-0.003	-0.001	-0.003	0.000
Village CES (median)	0.008	0.005	0.006	0.037	0.029	0.051

Low and high urban-rural gap are the means for the bottom quintile (15) and top quintile (15) regions ranked by the urban-rural gap in mean per capita expenditures in 2009. Panel A is based on equation 7 and panel B is based on equation 8, evaluated at utility corresponding to median group expenditure in rural areas. Rich bias is the difference between gains evaluated at 90th versus 10th percentile of group expenditure in rural areas. Panel C is based on the Feenstra (1994) applied at the region or village/block level (median across all bilateral village/block comparisons across sectors within region).