NBER WORKING PAPER SERIES

ESTIMATING FLEXIBLE INCOME PROCESSES FROM SUBJECTIVE EXPECTATIONS DATA: EVIDENCE FROM INDIA AND COLOMBIA

Manuel Arellano
Orazio Attanasio
Samuel Crossman
Víctor Sancibrián

Working Paper 32922
http://www.nber.org/papers/w32922

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2024

We thank Alberto Abadie, Aureo De Paula, Laura Liu, Wilbert van der Klaauw, Johannes Wohlfart and participants at several seminars for useful comments. An earlier version of this paper was presented as the Jacob Marschak Lecture at the 2021 Africa Meeting of the Econometric Society. The India survey was organised by Britta Augsburg for the evaluation of a microfinance program. Britta provided useful information on the data and the environment where they were collected. Yahu Cong and José M. Cueto provided excellent research assistance. We acknowledge financial support from the ESRC's Centre for the Microeconomic Analysis of Public Policy at the IFS (grant no. ES/T014334/1). Sancibrián acknowledges financial support from Grant PRE2022-000906 funded by MCIN/AEI/10.13039/501100011033 and by "ESF +", the Maria de Maeztu Unit of Excellence CEMFI MDM-2016-0684, funded by MCIN/AEI/10.13039/501100011033, Fundación Ramón Areces and CEMFI. The paper contents do not reflect the views of the UK Government or the Government Economic Service. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Estimating Flexible Income Processes from Subjective Expectations Data: Evidence from India and Colombia

Manuel Arellano, Orazio Attanasio, Samuel Crossman, and Víctor Sancibrián

NBER Working Paper No. 32922
September 2024
JEL No. C1, C42, J3, O12

ABSTRACT

We develop a methodology for modeling household income processes when subjective probabilistic assessments of future income are available. This allows us to flexibly estimate conditional *cdf*s directly using elicited individual subjective probabilities, and to obtain empirical measurements of subjective risk and persistence. We then use two longitudinal surveys collected in rural India and rural Colombia to explore the nature of income dynamics in those contexts. Our results suggest linear income processes are rejected in favor of more flexible versions in both cases; subjective income distributions feature heteroskedasticity, conditional skewness and nonlinear persistence.

Manuel Arellano
CEMFI
arellano@cemfi.es

Samuel Crossman
UK Government Economic Service
samuelcrossman@gmail.com

Orazio Attanasio
Department of Economics
Yale University
87 Trumbull Street
New Haven, CT 06511
and CEPR
and also NBER
orazio.attanasio@yale.edu

Víctor Sancibrián
CEMFI
victor.sancibrian@cemfi.edu.es

1 Introduction

Households allocate current income between consumption and savings taking into account the uncertainty about their future income *as they perceive it*. How persistent they perceive their future income flows to be, their dispersion or asymmetry are all subjective features of households' income uncertainty that critically impact their spending plans. Those perceptions and their heterogeneity across households are also key determinants of economy-wide consumption inequality.

Conventional approaches to identify the stochastic process of uncertain variables are *indirect*, relying on statistical models of the dynamics of the realized variables and/or models of choice.[1] Under rational expectations, income dynamics as perceived by households may coincide with the dynamics of realized income, but this needs not be the case. An alternative *direct* approach is to rely on subjective probabilistic expectation questions from surveys. In this paper, we develop a methodology for modeling household income processes using subjective expectations of future income. Our approach is flexible enough to assess the extent of nonlinear persistence and non-Gaussian distributional features in households' perceptions. We then take our methods to subjective expectations data elicited within two surveys that were conducted in Colombia and India. Learning about the nature of income uncertainty is particularly important in developing economies where there tends to be more volatility.

Subjective expectations have been around for a while. For a long time they were received with skepticism by some, but the current evidence is that individuals are able to respond to probabilistic questions about variables that matter to them in a meaningful way (Manski, 2004, 2018; Delavande, Giné, and McKenzie, 2011). Specifically, for developing countries a lot of progress has been made in understanding the implications of different methods of eliciting expectations (Attanasio, 2009; Delavande, 2023).

[1] See surveys of the literature on earnings dynamics in Meghir and Pistaferri (2011), Arellano (2014), and Altonji and Vidangos (2023).

Some of the early work on subjective income expectations is due to Dominitz and Manski (1997b). They used responses to the probability questions in their survey to fit respondent-specific parametric distributions, which they compared with those implied by the income processes used in Hall and Mishkin (1982). They found that subjective dispersion measures varied across households and were not proportional to subjective medians (see also Dominitz (1998, 2001)). Attanasio and Augsburg (2016) were the first to use the subjective expectations data in the survey of Indian households in combination with current income to estimate an income process.

We are also motivated by recent work on flexible income processes. A recent literature has uncovered significant non Gaussian nonlinearities in the dynamics of realized incomes (Arellano, Blundell, and Bonhomme, 2017; Guvenen, Karahan, Ozkan, and Song, 2021). These nonlinearities are potentially relevant for individual behavior and policy design, like saving choices (De Nardi and Paz-Pardo, 2020) or optimal income taxation (Golosov and Tsyvinski, 2015). It is of great interest to find out if these nonlinearities are also present in the subjective expectations of poor households in developing contexts.

Our first contribution is to show how to identify and estimate a standard (log) linear dynamic model for household income, with and without fixed effects, using data on subjective expectations and current income. Our approach is to map the model directly to individual subjective probabilities, and in particular to the *observed* log odd ratios, which we regard as noisy measures of the model counterparts, subject to an additive elicitation error. Fixed effects estimation of the model parameters is robust to un-modelled distributional heterogeneity of elicitation errors and, contrary to indirect approaches based on realized income, does not suffer from Nickell bias (Nickell, 1981). This is a convenient feature of subjective expectation models, since unobserved disturbances do not contain future shocks but only measurement errors in the elicited probabilities.

We use the log-linear model as starting point that conveys the main ideas of our approach. We then propose a generalized estimation framework to deal with nonlinear income processes with unobserved heterogeneity. We consider a sieve

approach with a sequence of flexibly parameterized predictive conditional distributions. Despite their generality, these distributions can be cast as static fixed effects models that can be estimated by within-group methods. We also explore extensions with more general patterns of unobserved heterogeneity where log odd ratios can vary differentially with individual effects. Our approach allows us to estimate subjective measures of risk and persistence that may differ across observed income levels, the size of income shocks, and individual effects.

In our empirical analysis, we use two waves from both the Colombian and Indian surveys, combining expectations with actual income data. In fact, the combination of the two is essential to our approach. In both surveys, income expectations were collected using Dominitz and Manski (1997a,b) elicitation method, alongside realized income and other indicators of the nature and sources of earnings. Respondents are asked to provide a relevant range of variation for their future income. Next, they are asked to report the probability that their future income will exceed each of three equally-spaced points within their selected range. These elicited probabilities are the individual-level outcomes in our models.

We reject the standard linear model in the data on subjective expectations from the two surveys in favor of more flexible models. Subjective income distributions exhibit nonlinear persistence, along with dispersion and skewness that vary with current income levels and unobserved heterogeneity. Interestingly, we find a negative association between conditional dispersion and current income, and between conditional skewness and current income.

Estimated persistence plummets for poorer households experiencing large positive shocks, but not for relatively affluent households experiencing negative shocks. Those findings for the perceived risks of households in developing economies are partially consistent with the results found in Arellano et al. (2017) for the realized incomes of US households from the Panel Study of Income Dynamics (PSID). Essentially, we find low persistence for large positive shocks at the bottom of the income distribution as they do, but not for large negative shocks at the top. In interpreting the results, we argue that the nonlinear persistence we find for the poorest households is consistent with a poverty trap interpretation. We also find that unobserved heterogeneity matters, and is composed of household specific and

village level factors. Households with large fixed effects have more persistent histories overall and less variability in persistence with current income and shock size. The pattern of nonlinear persistence is robust to allowing for more general forms of unobserved heterogeneity, although quantitatively its importance is reduced.

The paper is structured as follows. In the next section, we discuss how probabilistic subjective expectations are elicited in the two surveys that we use. Section 3 lays out our modeling and estimation framework, first for a linear process with fixed effects and then for more flexible models. Section 4 describes the two survey data sets that we use for the analysis. In Section 5, we present our empirical results. Finally, we conclude in Section 6. The appendices contain additional results and technical material.

2 Eliciting subjective expectations

Designing subjective expectation questionnaires to elicit information about respondents' perceived probability distribution of future variables is challenging. Several open questions remain in the growing literature on the topic, ranging from the establishment of a metric for the variables of interest to the way conditional and unconditional probability measures are elicited. As a consequence, important choices need to be made throughout the process.

In this section, we first briefly discuss some of the outstanding issues in the literature and then describe the approach used to elicit subjective expectations in our two surveys, which employed similar methods and questionnaire designs. While not particularly novel, it is useful to describe and relate them to possible alternatives.

2.1 Anchoring subjective expectations

A first issue in the design of subjective expectation questions is establishing an anchor and a metric for the variable whose probability distribution is being elicited.

In eliciting the probability distribution of future income, two different approaches have been used. In some surveys, the current value of income is used as an

implicit anchor, and respondents are asked the probability of a number of possible percentage changes of future income relative to current income. In other contexts, respondents are asked to provide a range of possible values, often the *minimum* and *maximum* for future income. These values are then used to define a number of intervals and respondents are asked the probability that future income will fall in each of these intervals. This approach was developed by Dominitz and Manski (1996, 1997a,b) and has been widely adopted in surveys across both developed and developing economies, including the Indian and Colombian datasets we use. The precise formulation of these questions is described in an elaborate script, which is reported in Appendix E.

Morgan and Henrion (1990) point out that asking first for the minimum and maximum of possible income realizations may help reduce two common problems in the elicitation of expectations. The first is *overconfidence*, wherein respondents focus too much on central tendencies and therefore understate the true uncertainty that they face — asking about the minimum and maximum first helps to prime respondents to think about the full range of probable realisations. The second is *anchoring* or *framing*, whereby figures provided by the interviewer might influence the responses provided: if the chosen *cdf* support points are specified by the interviewer, respondents may be inclined to think these points are salient for one reason or another, and therefore likely to restrict their answers around those values.

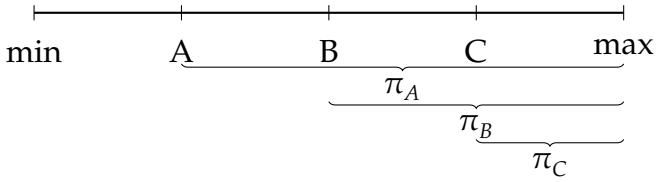
Using current income as an anchor for future income might be problematic when the former is unusually low or high, in that it is unclear whether the elicited probability distribution around that value is particularly informative. This is avoided by the minimum/maximum approach as respondents can *re-center* their answers around the information they have. On the other hand, it is not obvious whether the elicited minimum/maximum values are really what they are labeled to be or rather some arbitrary low or high percentile of the subjective future income distribution.

2.2 Eliciting subjective probability distributions

Having registered the minimum and maximum, the interviewers elicit from respondents information on several intermediate points on the subjective cumulative

distribution function (*cdf*) of future total household income. The minimum and maximum are used to compute, using a simple pre-specified algorithm, a set of J points within the support of the distribution of future income $\{y_{k1}, y_{k_2}, \ldots, y_{k_j}\}$. Respondents are then asked about the probabilities $\{\pi_{k1}, \pi_{k_2}, \ldots, \pi_{k_j}\}$ they assign to their next period income being larger than these points. In both survey data from India and Colombia, $J = 3$ was used. Furthermore, as shown in Figure 1, each sub-interval is of equal size.

FIGURE 1. A description of the elicitation process.



Note. The figure illustrates the way of eliciting three points on the subjective income distribution. Respondents are asked to provide $y_{\min}$ and $y_{\min}$. The interviewer then computes $y_B = (y_{\min} + y_{\max})/2$, $y_A = (y_{\min} + y_B)/2$ and $y_C = (y_B + y_{\max})/2$, and proceeds to elicit probabilities π_A, π_B and π_C.

A possible objection to this method is that respondents, especially from disadvantaged backgrounds, might be unfamiliar with the concept of probability and with its translation into numerical values. These issues might be particularly relevant in the context of developing countries, where respondents often have no or very limited formal education. In such a situation, the use of preliminary priming questions that could familiarize respondents with the theoretical constructs that researchers want to elicit, as well as the use of visual aids, may be advisable (Delavande et al., 2011).

For this reason, in both surveys, respondents were first *primed* in the use of the concept of probability and conditional probability through specific examples about future uncertain events. In particular, a sequence of questions was asked about the likelihood of rain, designed to ensure that probabilities are non-decreasing. For instance, the question "What is the probability that it will rain tomorrow?" is

followed by "What is the probability that it rains in the next seven days?", pointing out that the latter should be no smaller than the former.

As a form of visual aid for the probability questions, a *ruler* graded from 0 to 100 was used. Respondents were instructed to point to 0 to express the certainty that an event will not happen, to 100 for the certainty that it will, and to intermediate points to express uncertainty. While this approach seems to have worked in this context, it is not a silver bullet: different methods might be necessary in different contexts.[2]

During the data collection, interviewers were instructed to correct respondents' inconsistent answers during the training phase but not during the actual subjective expectations questions about future income. Respondents might, therefore, provide inconsistent answers, thereby flagging possible quality problems in the data. We explore this in depth in Section 4.

When designing expectations questions, an important choice is the number of intervals into which the range of values identified by the minimum and maximum is divided and the placement of cutoff points y_{k_j} for $j = 1, \ldots, J$. While a high number of cutoff points would increase the information on the *cdf*, improving the ability to fit flexible and possibly complex subjective *cdf*'s, such high values of J might impose an excessive burden on respondents and jeopardize the quality of the data. In a variety of contexts, J has been set at 1 or 3.

Another issue is whether the minimum and maximum should be treated as the genuine minimum and maximum of future income or as values where the *cdf* takes values relatively close to 0 or 1, respectively. We avoided committing to a specific interpretation of the minimum and maximum values, so that they play no role in our formal analysis beyond providing a range of respondent-specific points at which the *cdf* is elicited.

[2]Other surveys have asked respondents to allocate stones, balls or other items available in the local context, into a number of bins, see, for instance, Delavande and Rohwedder (2008).

3 Mapping income processes to subjective expectations

In this section, we show how to use data on subjective expectations to estimate models of household income dynamics, as perceived by respondents. We discuss the econometric approach we take to this problem, and show that the use of subjective expectations data poses inference problems that are conceptually different from those present when estimating dynamic models using actual income realisations. We start our discussion with a relatively simple (log) linear model, which is particularly useful in conveying the main ideas of our approach. We then generalize our approach and show that these data can also be used to estimate more complex and flexible income processes.

We interpret the models of the income processes as representing the conditional subjective probability distribution respondents hold, given the information available to them, including current income and other conditioning variables. To estimate the parameters of these models we then match the answers respondents give to the subjective expectations questions to the corresponding quantities implied by the statistical model we specify. As we discuss below, the identification of the structural parameters of the statistical models we consider relies on a number of assumptions. However, we argue that such assumptions are different and weaker than those used when estimating such models with actual income realisations. This approach allows us to use a wide class of estimators without incurring the biases that would affect such estimators when using actual income realizations.

3.1 Modeling approach

We take advantage of the availability of subjective expectations data to fit a model for the conditional *cdf* directly to the observed individual subjective probabilities. Let a household's subjective (conditional) cumulative probability distribution of log future income $y_{i,t+1}$ be denoted as

$$F_{it}(r) = P\left(y_{i,t+1} \leq r \middle| I_{it}\right), \tag{1}$$

where I_{it} denotes the information set available to household i in period t. As discussed in Section 2, the survey elicitation process employed in the datasets we use yields noisy measurements p_{jit} of $F_{it}\left(r_{jit}\right)$ for $r_{jit} = r_{it}^{\min} + \left(r_{it}^{\max} - r_{it}^{\min}\right) j/4$ $(j = 1, 2, 3)$, or equivalently of the subjective cumulative odds

$$\ell_{jit}^{*} = \text{logit}\left[F_{it}\left(r_{jit}\right)\right], \tag{2}$$

where $\text{logit}(p) = \ln\left[p/(1-p)\right]$. We model the log of the subjective cumulative odds, so that the outcomes we consider have an unlimited range of variation.[3] We also allow for a survey elicitation error ε_{jit}, which is plausibly assumed to be additive in the log of cumulative odds, so that the observed cumulative odds $\ell_{jit} = \text{logit}\left(p_{jit}\right)$ are given by

$$\ell_{jit} = \ell_{jit}^{*} + \varepsilon_{jit}. \tag{3}$$

We assume that ε_{jit} is a classical measurement error, in the sense that it is mean independent of I_{it}, the information set in equation (1). Apart from that, we allow for dependence in ε_{jit} across j and t. Moreover, the variance or other moments of ε_{jit} may change with j and t, and may also depend on variables in the information set. Modeling the distribution of elicitation errors is of separate interest, but the linearity assumption allows us to leave this distribution unmodeled while being robust to a variety of elicitation error configurations.

We only observe three points of F_{it} for each unit, but many different points across units. The general idea is to learn by combining data for all units; as long as there is sufficient variability in r_{jit} and common features in the probability distributions across units, they are potentially nonparametrically identifiable.

[3] Notice that this transformation rules out observations with $p_{jit} = 0$ or $p_{jit} = 1$. In Appendix D, we explore an alternative to (2) that introduces an adjustment to the logit function that can be interpreted as proportional to the accuracy of the elicitation process. This approach allows us to retain these observations and verify that our empirical results are largely unchanged.

Information set. The information set is assumed to be Markovian in the sense that given the current values of the relevant variables, values from earlier periods cannot reduce subjective prediction uncertainty. In our analysis, the Markov property is assumed to hold conditionally given unobserved heterogeneity.

The set I_{it} consists of time-varying and time-invariant characteristics. The time-varying variables include observable current income y_{it} and indicators x_{it} of the nature and sources of income, such as the number of earners in the household, as well as household demographics. As for the time-invariant characteristics, we adopt a latent variable approach, assuming that they can be captured by an unobservable individual effect α_i. This effect is intended to encompass both household-level characteristics and geographical (say, village-level) characteristics. We therefore assume that $I_{it} = (y_{it}, x_{it}, \alpha_i)$. The individual effect α_i may be correlated with $\left(r_{jit}, y_{it}, x_{it}\right)$.

A system of equations. The relationship between the information set available to individual households (part of which might be unobservable to the econometrician) and the elicited conditional *cdf*s depends on the specific model being considered. Here we define such a relationship as a function g, so to obtain:

$$\ell^*_{jit} = g\left(r_{jit}, y_{it}, x_{it}, \alpha_i\right) \qquad (i = 1, \ldots, n; j = 1, 2, 3; t = 1, 2) \tag{4}$$

where g is a non-decreasing function in its first argument. We specify below the function g corresponding to different models of income dynamics. Thus, our econometric model consists of a system of six equations for n households with the addition of measurement errors in elicited probabilities.[4]

Identification of nonlinear panel data models with continuous outcomes and unobserved heterogeneity has been discussed in Evdokimov (2010), Arellano and Bonhomme (2016), Hu (2017), and Schennach (2022), amongst others. Nonparametric identification of the response function g and the conditional distribution of α_i in model (3) and (4) can be established using the arguments in Evdokimov

[4]Note that an additive α_i could be reflecting both persistent elicitation differences (that is, part of measurement error) or heterogeneity in income risk. This distinction will matter for interpretation when documenting the effect of "heterogeneity" in nonlinear models.

(2010), under the assumption that the additively separable disturbance terms are conditionally independent over time, and independent of the individual effect α_i.

Next, we consider alternative specifications of the income model, starting with the simplest version, which assumes linearity and corresponds to models that have been widely used in the literature on income processes.

Income processes. In a life-cycle model of income and consumption choices, a popular specification decomposes household income into the product of a deterministic (or profile) component, which might include a fixed effect, and a stochastic component, often in the form of persistent shocks with autoregressive dynamics (sometimes also including transitory shocks). A log-linear model of this kind with no transitory shocks can be written as

$$Y_{i,t+1} = Y_{it}^{\rho} V_{i,t+1} \exp\left(p_{i,t+1} + \alpha_i\right),$$

where Y_{it} is the level of income for household i at time t, $V_{i,t}$ innovations to income, $p_{i,t}$ captures household age and demographic variables, and α_i represents the fixed effect. In both of our data sets, these fixed effects can be decomposed into village-level and purely idiosyncratic effects. We omit this distinction here for notational simplicity. Taking logs, we have

$$y_{i,t+1} = \rho y_{it} + p_{i,t+1} + \alpha_i + v_{i,t+1}, \tag{5}$$

where $y_{it} = \ln Y_{it}$ and $v_{it} = \ln V_{it}$.

One could consider decomposing the stochastic part of income into persistent and transitory components. In a standard persistent/transitory model, consumers are assumed to observe the values of the two components as separate state variables, whereas they remain unobserved to the modeler. However, when conditioning on current income as we do, this situation introduces a measurement-error problem that can be dealt with instrumental variables. Such an approach cannot be used in a two-wave panel like ours. While we do not consider this possibility in our application, we do estimate multiple-state processes that include indicators of the nature and sources of income, whose motivation is not entirely different from that behind unobservable income component models.

3.2 Predictive distributions for linear income processes

We first illustrate how our approach allows us to estimate conditional distributions of subjective income risk when the underlying income process is a standard log-linear autoregressive model. This is a convenient benchmark which provides a simple framework to illustrate identification and estimation issues, while highlighting the benefits of using subjective expectations data relative to a more standard approach using income realizations.

Considering a first-order autoregressive process with fixed effects,[5] we can rewrite equation (5) as follows:

$$y_{i,t+1} = \alpha_i + \rho y_{it} + \sigma v_{i,t+1}, \tag{6}$$

where $v_{i,t+1}$ are assumed to have a logistic distribution independent of y_{it} and α_i. The corresponding conditional *cdf* is then

$$\begin{aligned} P\left(y_{i,t+1} \leq r \middle| y_{it}, \alpha_i\right) &= P\left(v_{i,t+1} \leq \frac{r - \alpha_i - \rho y_{it}}{\sigma} \middle| y_{it}, \alpha_i\right) \\ &= \Lambda\left(\frac{r - \alpha_i - \rho y_{it}}{\sigma}\right), \end{aligned}$$

where $\Lambda(x) = (1 + \exp(-x))^{-1}$ is the standard logistic *cdf*. Applying the logit transformation, it follows that in this case g in (4) is linear, since we can write

$$\ell_{jit} = \ell^*_{jit} + \varepsilon_{jit} = \beta_0 r_{jit} + \beta_1 y_{it} + \eta_i + \varepsilon_{jit}, \tag{7}$$

where $\beta_0 = 1/\sigma$, $\beta_1 = -\rho/\sigma$ and $\eta_i = -\alpha_i/\sigma$. The logit transformation in (2), therefore, allows us to map the "structural" parameters ρ and σ to the "reduced-form" estimation parameters in equation (7).[6] Equation (7) is a linear panel model with fixed effects and strictly exogenous regressors, and a standard within group estimator yields consistent estimates of the parameters.

[5] We discuss specifications with time-varying characteristics below; see Section 3.3. On a similar note, we include time (wave) effects in all models that we consider, but omit them from explicit formulas for notational simplicity.

[6] We would obtain a similar mapping if we assumed, for instance, that $\ell_{jit} = \text{probit}\left(p_{jit}\right)$ and $v_{i,t+1} \sim \text{N}(0,1)$, independent of y_{it} and α_i.

As discussed in Section 4 below, in both surveys we use, the observations are clustered in villages. Therefore, when considering fixed effects, we allow for village-specific means via the following decomposition:

$$\eta_i = \bar{\eta}_{v(i)} + \tilde{\eta}_{i,v(i)},$$

where the subscript $v(i)$ indicates the village of household i, $\bar{\eta}_{v(i)}$ is a village-specific mean, and $\tilde{\eta}_{i,v(i)}$ denotes the deviation of the individual fixed effect from the village average. Regardless of whether the variance of the purely idiosyncratic fixed effects is constant across villages or not, we can estimate the variance of the village fixed effects and the unconditional variance of the purely idiosyncratic fixed effects.

Subjective expectations and income realizations. Despite superficial similarities there are profound differences between the subjective expectation and observed income approaches. First, with subjective expectations data, an AR(1) model without fixed effects can be estimated on a *single cross-section*, as information on *expected future income* (on the left-hand side) is provided by subjective expectations. If fixed-effects are included, the variance of the shock (a measure of risk) can still be estimated on a single cross-section, and the full model would require two waves of data — whereas the observational approach would need at least three.

Second, estimates from subjective expectations represent the perceptions individual households have of their own income, even if they do not have rational expectations. Such an object is what is relevant for household consumption and saving decisions.

Finally, estimation of the model using subjective expectations data does not suffer from the so-called Nickell bias, and so there is no need to use instrumental variable techniques, despite the small time dimension. This is typically not the case when using only income realizations. The reason is that outcomes are not future incomes but rather points in the predictive distribution; therefore, the error term does not contain future shocks but only measurement error in predictive probabilities.

While the linear model is useful to illustrate how subjective expectations can be used to recover the parameters of a standard income process, it still imposes a number of tight restrictions regardless of whether subjective expectations or realized

income data are used in estimation. For example, persistence ρ and dispersion σ are common to all households in equation (7), a restriction that we relax next.

3.3 Enlarging the state space

The existing literature has mainly focused on single-state processes in which current income (or a persistent/transitory decomposition of income) is a sufficient statistic for the information set in a household's predictive distribution of future income. However, it is possible that indicators of the nature and sources of income and/or the occurrence of specific shocks help predict future income over and above total current income. If so, consumption decisions might depend on the joint probability distribution of a vector of future variables. Multivariate models of income dynamics are beyond the scope of this paper, but it is still of interest to find out if our subjective probabilities of future income depend on a larger state space than current income.

Thus, additional flexibility can be added by including relevant time-varying household characteristics x_{it} in the conditioning set, which extends (7) to

$$\ell_{jit} = \beta_0 r_{jit} + \beta_1 y_{it} + \delta_0' x_{it} + \delta_1' x_{it} y_{it} + \eta_i + \varepsilon_{jit}. \tag{8}$$

In our empirical analysis, we also provide results for models of this type.

3.4 Flexible income processes

We now generalize the linear model in equation (7) to the following specification:

$$\ell_{jit} = \beta_0(r_{jit}) + \beta_1(r_{jit})\psi(y_{it}) + \beta_2(r_{jit})\eta_i + \varepsilon_{jit}, \tag{9}$$

where $\beta_s(r_{jit})$ for $s = \{0, 1, 2\}$ and $\psi(y_{it})$ are functions such as splines or orthogonal polynomials and, again, we omit time-varying observables x_{it} for simplicity. Models with additive fixed effects correspond to setting $\beta_2(r_{jit}) = 1$, whereas the full

generality of (9) allows for interactive effects.[7] The linear model (7) is a special case of (9) with linear $\beta_0(\cdot)$ and $\psi(\cdot)$ and constant $\beta_1(\cdot)$ and $\beta_2(\cdot)$.

This model is reminiscent of distribution regression, but the empirical setup is rather different. In distribution regression, one would use realized data on $y_{i,t+1}$ and estimate a sequence of logit or probit regressions for binary outcomes defined as $\mathcal{I}(y_{it} < r)$ to get estimates of $\beta_k(r)$ for different chosen values of r.[8] In our context, we observe $P\left(y_{i,t+1} \leq r \middle| y_{it}, \alpha_i\right)$ for $r = r_{jit}$, so that we can fit these observed probabilities to the specific model we consider. To perform such an exercise, the functions $\beta_s(r)$ and $\psi(\cdot)$ need to be parameterized. Implementation and estimation details are discussed in Section 3.5 below.

Measuring dispersion, skewness, and persistence. The coefficients of the splines and polynomials in equation (9) may not have a straightforward or meaningful interpretation on their own. Instead, we use them to compute quantile-based measures of dispersion, skewness and persistence, which characterize some of the properties of the nonlinear models of interest. To do so, we need to calculate the implied quantiles from our conditional *cdf* model. Let $q_{it}(\tau)$ be the τ quantile from the model for some $\tau \in (0,1)$, which is the value of r that solves the equation

$$g(r, y_{it}, x_{it}, \alpha_i) = \text{logit}(\tau). \tag{10}$$

For example, for the linear autoregressive model in equation (6), the conditional quantile is defined as

$$q_{it}(\tau) = \rho y_{it} + \alpha_i + \sigma \, \text{logit}(\tau). \tag{11}$$

More generally, the solution can be found numerically using bracketing or interpolation methods.

[7] In principle, interactions between η_i and y_{it} could add even greater flexibility, but we did not explicitly include them to preserve the simplicity of estimation given the characteristics of our samples. Still, the growth-rate form of the model that we discuss in Section 3.5 effectively incorporates such interactions.

[8] See Foresi and Peracchi (1995), and Chernozhukov, Fernández-Val, and Melly (2013).

A standard measure of dispersion is the interquantile range:

$$\text{IR}_{it}(\tau) = q_{it}(\tau) - q_{it}(1-\tau), \tag{12}$$

where usually $\tau = 0.75$ or $\tau = 0.90$. For example, for the linear AR(1) model we have

$$\text{IR}_{it}(\tau) = \sigma \times 2\,\text{logit}(\tau).$$

Dependence of IR on y_{it} and/or η_i indicates heteroskedasticity. Similarly, the Bowley-Kelley measure of skewness for some $\tau > 0.5$ is given by:

$$SK_{it}(\tau) = \frac{[q_{it}(\tau) - q_{it}(0.5)] - [q_{it}(0.5) - q_{it}(1-\tau)]}{q_{it}(\tau) - q_{it}(1-\tau)}. \tag{13}$$

Finally, in nonlinear models we use the measure of persistence proposed in Arellano et al. (2017), which is defined as :

$$\rho_{it}(\tau) = \frac{\partial q_{it}(\tau)}{\partial y_{it}}. \tag{14}$$

Using the chain rule, $\rho_{it}(\tau)$ can be written as a scaled derivative effect of realized income in our model for the cumulative distribution:

$$\rho_{it}(\tau) = -\frac{\partial g(q_{it}(\tau), y_{it}, x_{it}, \alpha_i)}{\partial y_{it}} \Bigg/ \frac{\partial g(q_{it}(\tau), y_{it}, x_{it}, \alpha_i)}{\partial r}. \tag{15}$$

For the linear AR(1) model we simply have $\rho_{it}(\tau) = \rho$. In general, the persistence of the process will depend on the position of a household in the distribution of current income, fixed effects, and the value of τ.

Note that an equation such as (6) relates the *realized* shock $v_{i,t+1}$ with rank $\Lambda(v_{i,t+1})$ to the *realized* outcome $y_{i,t+1}$ given $y_{i,t}$. However, we can also consider hypothetical shocks and their corresponding hypothetical outcomes. For example, we can ask what would be the $t+1$ outcome if the $t+1$ shock was one with rank $\tau \in (0,1)$. This is precisely the information provided by the conditional quantile function (11). In the nonlinear generalization, the persistence measure (14) provides the weight of current income in the function that produces future income when a household is hit by a shock of rank τ.

In our analysis we do not rely on realized future outcomes and realized future shocks, but on the subjective conditional probability distribution of future outcomes, which allows us to speak about the impact of *potential (subjective) shocks* on potential future outcomes.

3.5 Implementation and estimation

We now discuss the specification of the various functions that enter the flexible income model in equation (9) and the estimation of the relevant parameters.

3.5.1 Specification

The functions $\beta(\cdot)$ and $\psi(\cdot)$ in (9) need to be parameterized. We first reformulate the model we are considering as a predictive distribution for income growth, since departures from linearity may be better captured for income changes than for levels. This change is immaterial for the linear model, but leads to different approximating models for nonlinear specifications.

Predictive distributions for growth rates. The elicited probabilites p_{jit}, which are noisy measures of $F_{it}\left(r_{jit}\right) = P\left(y_{i,t+1} < r_{jit} \middle| I_{it}\right)$, also measure $F_{it}^{\Delta}\left(s_{jit}\right) = P\left(\Delta y_{i,t+1} < s_{jit} \middle| I_{it}\right)$ for $s_{jit} = r_{jit} - y_{it}$. This is so because y_{it} is part of the information set. The function $F_{it}^{\Delta}\left(s_{jit}\right)$ is the predictive distribution of future income growth and is connected to $F_{it}\left(r_{jit}\right)$ by a simple translation of its argument: $F_{it}(r) = F_{it}^{\Delta}(r - y_{it})$. Thus, in a non-parametric sense, there is no difference between estimating one function or the other. However, in practice it may be better to estimate flexible models for $F_{it}^{\Delta}(s_{it})$ instead of $F_{it}(r)$, even if the interest is in $F_{it}(r)$.

If the true process is a random walk, $F_{it}^{\Delta}(s)$ will be a constant *cdf*, which does not depend on y_{it}, so that modeling $F_{it}^{\Delta}(s)$ is equivalent to modeling departures from a random walk. More generally, it can be expected that standardizing the range of variation in the argument by subtracting y_{it} will help modeling. Another consideration is that in the *cdf* of $\Delta y_{i,t+1}$, the nonlinearities considered in Arellano

et al. (2017) are close to tail departures from linearity in a single-index logit or probit, but require translations of the index in the *cdf* of $y_{i,t+1}$.[9]

In the linear case, the model to be estimated remains essentially unchanged by targeting log income changes instead of log levels.[10] However, for nonlinear specifications, the actual flexible model to be estimated will be different:

$$\ell_{jit} = \beta_0^\dagger(s_{jit}) + \beta_1^\dagger(s_{jit})\psi\left(y_{it}\right) + \beta_2^\dagger(s_{jit})\eta_i + \varepsilon_{jit}. \tag{16}$$

For a given level of complexity, the functions $\beta_k^\dagger(s_{jit})$ may be better approximators to a class of models of interest than $\beta_k(r_{jit})$.

Implementation. Our empirical specification will be based on (16), taking the components of $\psi(\cdot)$ in a polynomial basis of functions. We specified $\psi(\cdot)$ as a vector of low-order Hermite polynomials in standardized current income. The functional coefficients $\beta_k^\dagger(\cdot)$ are taken as natural cubic splines on s_{jit}, also entering in standardized form in those functions. In general, fitting a natural cubic spline with $L \geq 2$ knots requires estimating L parameters. Further details are provided in Appendix B.

Note that a flexible specification of the $\beta_k^\dagger(\cdot)$ coefficients can undo the possible restrictiveness of the logistic transformation that we use. For example, if the income process is a random walk with non-logistic shocks, for a sufficiently flexible specification of the intercept term $\beta_0^\dagger(\cdot)$, the formulation $P\left(\Delta y_{i,t+1} < s_{jit} \middle| I_{it}\right) = \Lambda(\beta_0^\dagger(s_{jit}))$ will capture a broad class of *cdf*s regardless of $\Lambda(\cdot)$.

---

[9] This observation can be made precise using the simple switching income process with nonlinear persistence in Arellano et al. (2017, equations (S6) and (S7)), where the predictive probit of income growth for a household around median income is a straight line independent of income. For a low (high) income household, the line jumps upwards (downwards) at right (left) tail values of income changes, but remains a straight line for most of the range of variation. In contrast, the predictive probit of log income will change with the income level across the entire income distribution, compounded with additional nonlinear variation in the tails.

[10] The linear reparameterized equation is

$$\ell_{jit} = \beta_0^\dagger s_{jit} + \beta_1^\dagger y_{it} + \eta_i + \varepsilon_{jit},$$

where $\beta_0^\dagger = \beta_0$, $\beta_1^\dagger = \beta_1 + \beta_0 = (1-\rho)/\sigma$ and $s_{jit} = r_j - y_{it}$.

3.5.2 Estimation

In specifications with $\beta_2^\dagger(s_{jit}) = 1$, the model is a *static* fixed effects regression that can be consistently estimated using the within-group estimator. Our estimation approach allows for the introduction of a ridge penalty $\lambda > 0$ on the higher-order coefficients of the spline to control overfitting in the more flexible specifications, although the results reported in the paper set $\lambda = 0$.

Rearrangement. Since monotonicity of g is not imposed, the estimated curve may be non-monotone. To address this issue we follow the method proposed in Chernozhukov, Fernández-Val, and Galichon (2010), which consists in sorting the original estimated curve into a monotone rearranged curve.

Substantial violations of monotonicity may signal misspecification. When using flexible specifications in growth-rate form, the rearranged and non-rearranged estimated probability distribution functions that we obtain are virtually identical.

Estimating models with interacted fixed effects. In specifications where $\beta_2^\dagger(s_{jit})$ depends on unknown parameters, there is an incidental parameters problem in the fixed-effects approach. Specifically, the least-squares estimator of the model's common parameters based on their joint estimation with $(\eta_1, \ldots, \eta_n)$ suffers from an errors-in-variables bias and is not consistent in a short panel. The difficulty owes to the fact that $\hat{\eta}_i$, which is used as a regressor, is a noisy estimator of η_i.

To obtain consistent estimates that take into account small-T errors in the estimated fixed effects, one can resort to either method-of-moments or pseudo maximum-likelihood approaches. A method-of-moments approach to estimating a random coefficients model for panel data is developed in Chamberlain (1992). Here we use a simple extension of the linear model in (7) to illustrate how to construct a linear instrumental-variable estimator of the kind that we employ in obtaining our empirical results, and defer to Appendix B a more general discussion of the estimation of models with interacted effects.

A simple instrumental-variable estimator. Let us consider the model

$$\ell_{jit} = \beta_0 r_{jit} + \beta_1 y_{it} + \left(1 + \beta_2 r_{jit}\right)\eta_i + \varepsilon_{jit}, \tag{17}$$

which boils down to the standard linear income process when $\beta_2 = 0$. However, solving for the conditional quantile function, we can see that this model corresponds to a very different income process with heterogeneous risk and persistence that generalizes equation (6) to

$$y_{i,t+1} = -\frac{\eta_i}{\beta_0 + \beta_2\eta_i} - \frac{\beta_1}{\beta_0 + \beta_2\eta_i} y_{it} + \frac{1}{\beta_0 + \beta_2\eta_i} v_{i,t+1}.$$

To get an estimating equation, first note that taking deviations from individual means does not remove unobserved heterogeneity from model (17):

$$\tilde{\ell}_{jit} = \beta_0 \tilde{r}_{jit} + \beta_1 \tilde{y}_{it} + \beta_2 \tilde{r}_{jit}\eta_i + \tilde{\varepsilon}_{jit}, \tag{18}$$

where $\tilde{\ell}_{jit} = \ell_{jit} - \bar{\ell}_i$, $\tilde{r}_{jit} = r_{jit} - \bar{r}_i$, and so on. However, we can use the transformation

$$r_{jit}\bar{\ell}_i - \bar{r}_i\ell_{jit} = \beta_1\left(r_{jit}\bar{y}_i - \bar{r}_i y_{it}\right) + \tilde{r}_{jit}\eta_i + \left(r_{jit}\bar{\varepsilon}_i - \bar{r}_i\varepsilon_{jit}\right)$$

to substitute out $\tilde{r}_{jit}\eta_i$ in (18) and obtain

$$\tilde{\ell}_{jit} = \beta_0 \tilde{r}_{jit} + \beta_1 \tilde{y}_{it} + \gamma\left(r_{jit}\bar{y}_i - \bar{r}_i y_{it}\right) + \beta_2\left(r_{jit}\bar{\ell}_i - \bar{r}_i\ell_{jit}\right) + \xi_{jit}, \tag{19}$$

where $\xi_{jit} = (1 + \bar{r}_i)\,\varepsilon_{jit} - (1 + r_{jit})\bar{\varepsilon}_i$ and $\gamma = -\beta_1\beta_2$. Whereas the error term ξ_{jit} is mean independent of r_{jit} and y_{it} for all (t, j) (which we collect into w_i), $r_{jit}\bar{\ell}_i - \bar{r}_i\ell_{jit}$ is an endogenous variable in equation (19). To motivate an IV estimator, note that

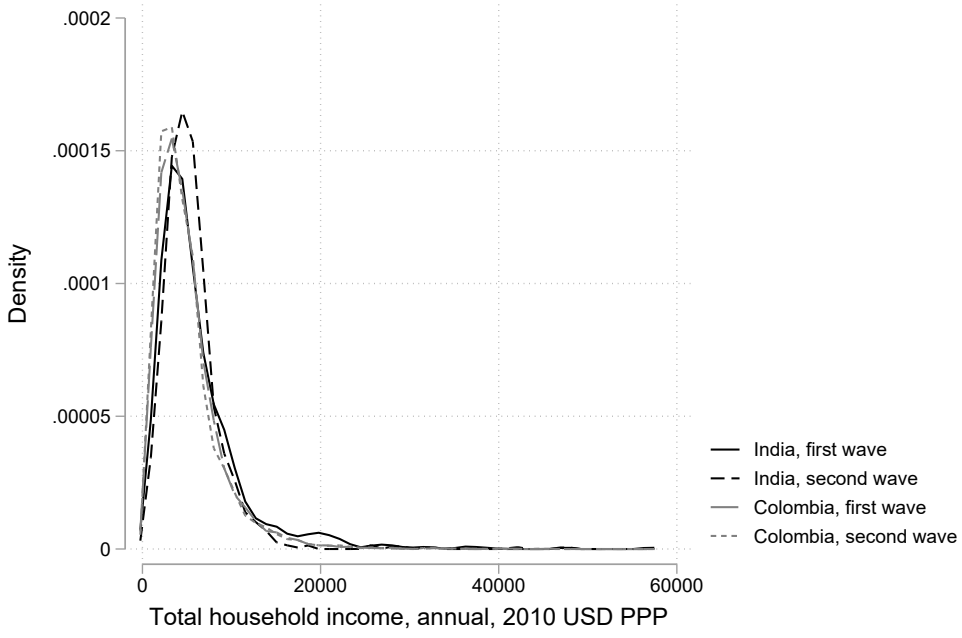
$$E\left[r_{jit}\bar{\ell}_i - \bar{r}_i\ell_{jit}\middle|w_i\right] = \beta_1\left(r_{jit}\bar{y}_i - \bar{r}y_{it}\right) + \tilde{r}_{jit}E\left[\eta_i\middle|w_i\right].$$

Approximating $E\left[\eta_i\middle|w_i\right]$ by the projection of η_i on $\bar{w}_i$ suggests using $\tilde{r}_{jit}\bar{r}_i$ and $\tilde{r}_{jit}\bar{y}_i$ as external instruments for $r_{jit}\bar{\ell}_i - \bar{r}_i\ell_{jit}$ and estimating equation (19) by two-stage least squares (TSLS). The restriction $\gamma = -\beta_1\beta_2$ is not required for identification and might be ignored to avoid nonlinear estimation.

4 Data

We use data on subjective income expectations in combination with data on realized income from two developing country contexts — rural India and Colombia. In both cases, the subjective income expectations were collected as part of broad surveys aimed at evaluating development interventions. Both interventions were targeted at a poor and rural population.

In Figure 2, we plot the distribution of reported total household income across countries and survey waves. Income is measured in 2010 PPP USD, which we use for both countries throughout the analysis. While the contexts we are studying are very distinct, the two distributions are remarkably similar, indicating that we are concerned with comparably poor populations. Average annual household income in the sample is $5,924 for India and $5,013 for Colombia, with a standard deviation of $4,632 and $3,759, respectively.



Note. The Figure shows the distribution of total household income in the two study populations, in 2010 PPP USD. Monthly income in Colombia is annualized for comparability.

FIGURE 2. Household income across study populations.

In what follows, we briefly describe the survey contexts and the characteristics of the respondents and their households. We then provide some evidence about the validity of the expectations data. In both surveys, subjective expectations data were elicited using the approach described in Section 2. The main difference between the two surveys is in the horizon of future income: in India future income refers to the following year, while in Colombia is the following month.

4.1 India

The data in India were collected in 64 villages in Anantapur, a district located in the southern state of Andhra Pradesh, for the evaluation of a microfinance intervention (loans for cow or buffalo); see Augsburg (2009) for additional details. A typical household we consider has five members and a male, 45 years-old household head. Most households belong to the "Other Backward caste", a collective term used by the Government of India to classify castes which are educationally or socially disadvantaged. A further 13% belong to the Scheduled Castes, 5% to the Scheduled Tribes, and the remaining 28% to the General Caste. More than 60% of household heads had not undergone any formal education, and only 10% had some primary education. The average household depends on three income sources, with agriculture being the primary activity — as farmers (25%) or as agricultural labourers (64%). Additional details can be found in Attanasio and Augsburg (2016, Table 1).

In Table 1, we present descriptive information on income sources and shocks, which we later integrate in our models and which provide contextual information on the importance of different sources of risk that households face. Households with less than three, three, and four or more income sources account for 42.3%, 37.2%, and 20.5% of the sample, respectively. For each such category, we report the percentage of income from farm-related activities (which includes agriculture) and the types of shocks experienced. Households with at most two income sources are relatively more likely to report no income from farm-related activities. Moreover, the likelihood of reporting no shocks is about 10%, health shocks about 20%, and agricultural shocks about 50-60%, quite uniformly across household categories.

TABLE 1. India – income shocks and sources

	≤2 sources	3 sources	4+ sources
Proportion	0.423	0.372	0.205
0% farm	0.347	0.155	0.054
Up to 50% farm	0.235	0.431	0.502
More than 50% farm	0.419	0.414	0.444
	1.00	1.00	1.00
No shocks	0.113	0.087	0.092
Health shock	0.233	0.194	0.200
Agriculture shock	0.531	0.606	0.603
Other shocks	0.123	0.113	0.105
	1.00	1.00	1.00

Note. The table shows relative frequencies (proportions) of different components of total household income for the Indian data, pooling across the two waves. The first row displays the proportion of households reporting up to two different income sources, three income sources, and more, respectively. The next three rows report the proportion of current income stemming from farm-related activities for each subgroup by income source. The final four rows show the relative frequency of the most important types of (negative) shocks faced by households during the previous year, again for each income source subgroup.

After the household baseline sample was interviewed in January/February 2008, a follow-up survey was conducted in April/June 2009. Respondents were asked to provide information on income and subjective expectations in both survey rounds. Of the 1,036 households that made the original sample, 947 were re-interviewed in the second wave. We drop observations with missing income or at least one reported probability and those with elicited expectations that violate basic probability laws, following the analysis in Attanasio and Augsburg (2016). This yields a balanced panel with $N = 770$ households. Details are reported in Table A.1 in Appendix A.1.[11]

[11]This attrition rate is slightly higher than that reported in Attanasio and Augsburg (2016). As documented in the Appendix, these differences are mostly due to removing outliers in reported and expected income.

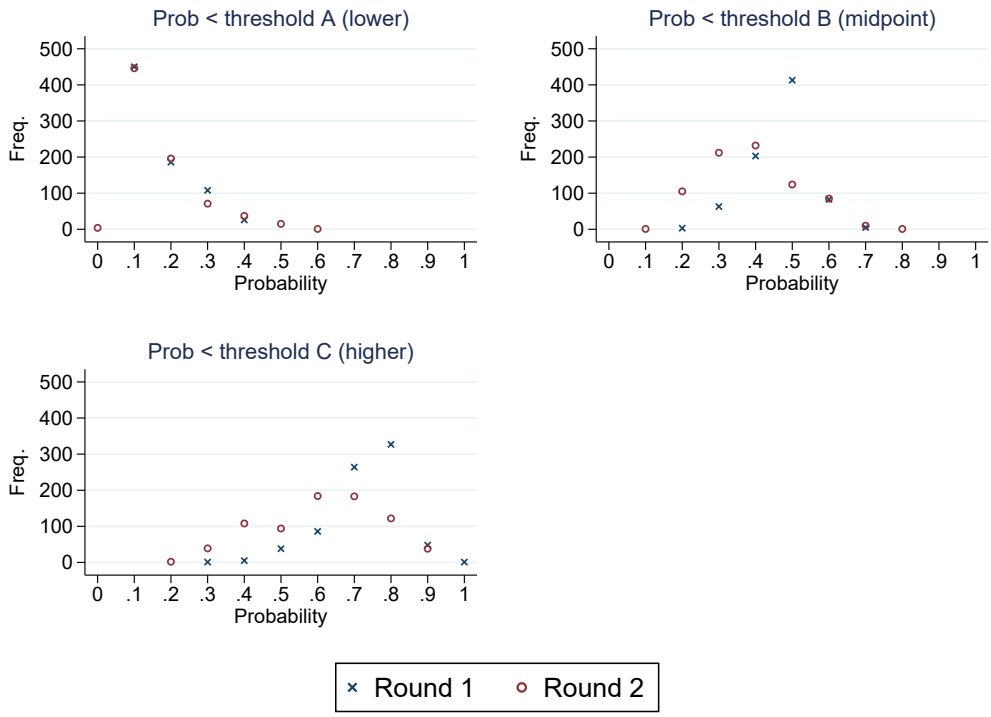
About a quarter of these households were clients of a microfinance institution (MFI) and had in 2008 loans provided livestock investment. The remaining households were either residing in the same villages or in villages the MFI considered targeting in the future. The data was collected to evaluate the provision of these livestock loans, which aimed at enabling households to engage in milk-selling as an additional income-generating activity, thereby reducing their dependence on outcomes of the main cropping seasons.

In both survey waves, the interviewers — who visited the respondents in their homes — elicited information on points on the respondents' subjective household income distribution. The technique discussed in Section 2 was used after explaining the approach in detail, practicing with rainfall questions, and using a ruler as a visual aid. Respondents were asked about their expected household income for the year following the interview. This interval was chosen considering the irregularity of income and to ensure key income periods were covered.

As discussed in detail in Attanasio and Augsburg (2016), respondents were not only willing to provide (expected) income information, but also provided sensible answers that reflected their beliefs. In particular, (i) over 97% of respondents provided responses to all three thresholds, (ii) violations of basic probability laws (monotonicity and wrong "direction") make up less than 1% of the sample, (iii) very few households bunch at 100% for the highest threshold or 0% for the lowest, indicating that the minimum and maximum expected income are well elicited, (iv) respondents made otherwise use of the entire range, although some bunching at multiple of 5s was observed (possibly because these were indicated on the ruler), and (v) expectations correlate sensibly with household characteristics. We report further details in Appendix A.1.

Figure 3 provides a summary of the distribution of elicited probabilities by threshold and survey round. The upper left panel displays the absolute frequencies for reported cumulative probabilities below the first threshold, which as expected are skewed to the right — the mode being at 0.1 in both survey rounds. On a similar vein, the distribution is mostly concentrated within the 0.3-0.6 range for the midpoint threshold, and skewed to the left for the highest threshold.

FIGURE 3. India: frequencies by threshold



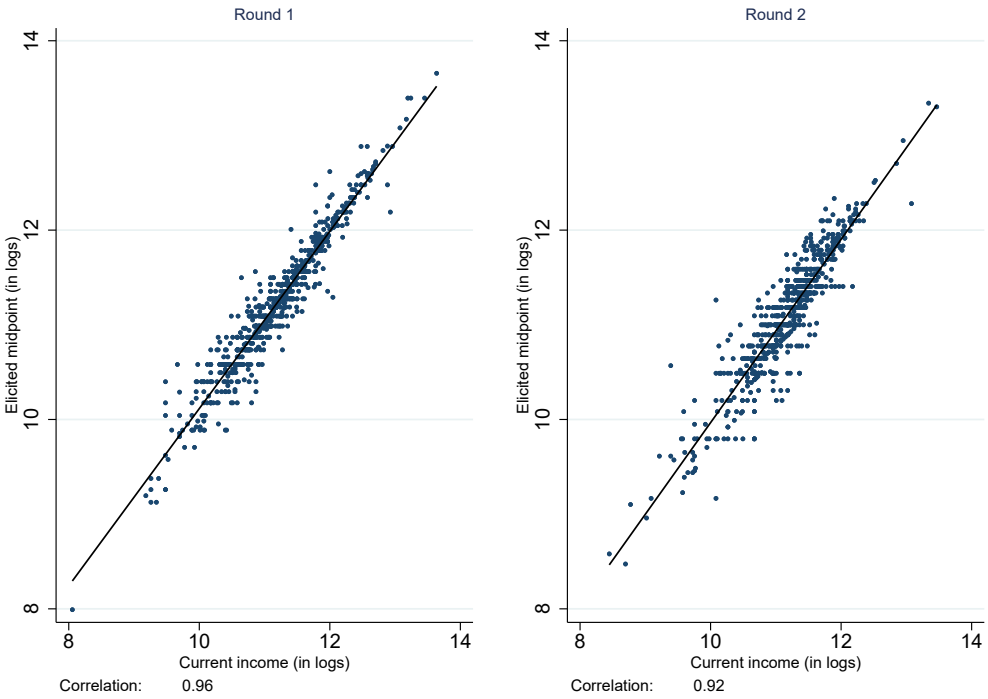
Note. The figure displays subjective probability frequency plots for each threshold and survey round, as shown in the legend. Probabilities are rounded to the nearest tenth.

Figure 4 plots together the elicited midpoint of the predictive distribution and realized current income, and shows an extremely high cross-sectional correlation between household's current income and their median subjective assessment for next year's income. This evidence suggests that the subjective income expectations data provide useful and meaningful information.

4.2 Colombia

The data in Colombia were collected in 122 of the country's poorest municipalities located in 26 of 34 departments to evaluate the introduction of a Conditional Cash Transfer (CCT) program, called *Familias en Acción* (FEA), a welfare program run by the Colombian government to foster the accumulation of human capital through

FIGURE 4. India: current income and reported midpoint



Note. The solid black corresponds to the linear regression fit.

improved nutrition, health, and education in rural Colombia. As many CCTs around the world, FEA pursued its objective through a cash transfer conditional on child vaccinations, development checks, school attendance, and courses for the mother. The program was targeted to the poorest sectors of society; recipients typically fall into the bottom 20% of Colombian households living in rural areas.[12]

The evaluation first conducted a baseline survey in 2002, approaching 11,500 and interviewing 11,462 households. We use data from the two follow-up survey rounds, conducted from July to November 2003 and again from November 2005 to

[12]In particular, recipients (or potential recipients, in the case of the evaluation sample) were in the lowest category of the SISBEN indicator, which is used to target most social programs and to set utility prices. See Attanasio, Battistin, and Mesnard (2012, Section 2).

March 2006, completing interviews to 10,743 and 9,463 households, respectively.[13] At the time of the second survey, about half of respondent households were target beneficiaries of *Familias en Acción*.

TABLE 2. Colombia – income shocks and providers

	1 earner	2 earners	3+ earners
Proportion	0.380	0.413	0.207
Up to 75% regular	0.145	0.269	0.223
More than 75% regular	0.080	0.359	0.434
100% regular	0.775	0.372	0.344
	1.00	1.00	1.00
No shocks	0.773	0.749	0.712
Health shock	0.089	0.093	0.124
Other shocks	0.138	0.158	0.163
	1.00	1.00	1.00

Note. The table shows relative frequencies (proportions) of different components of total household income for the Colombian data, pooling across the two waves. The first row displays the proportion of households with one earner, two earners or three or more earners (during the previous month). The next three rows report the proportion of current income stemming from regular sources (as opposed to occasional) for each subgroup by number of earners. Note that "more than 75% regular" excludes 75% and 100%. The final three rows show the relative frequency of the most important types of (negative) shocks faced by households during the previous year, again for each category of number of earners. Note that around 2.5% of households report having suffered both health- and non health-related shocks, which implies that the absolute frequencies for these two categories are slightly larger than reported above, for each category of number of earners.

Survey respondents are predominantly female (65%). Just over half (54%) are household heads, living in households that, on average, included another five members, with an average age of 43 years and with low levels of education: 24% of household heads have less than primary education, with an average of 3.5 years of schooling. Income is predominantly earned in the form of labour income, where most individuals tend to be informally employed (93%). About half of the working

[13]These figures correspond to the first row in Table A.3 in Appendix A.2, which also provides additional information on response rates and sample sizes.

individuals in the sample work in agriculture, while others, for example, work as domestic servants. The survey design and context are described in detail in Attanasio et al. (2012).[14]

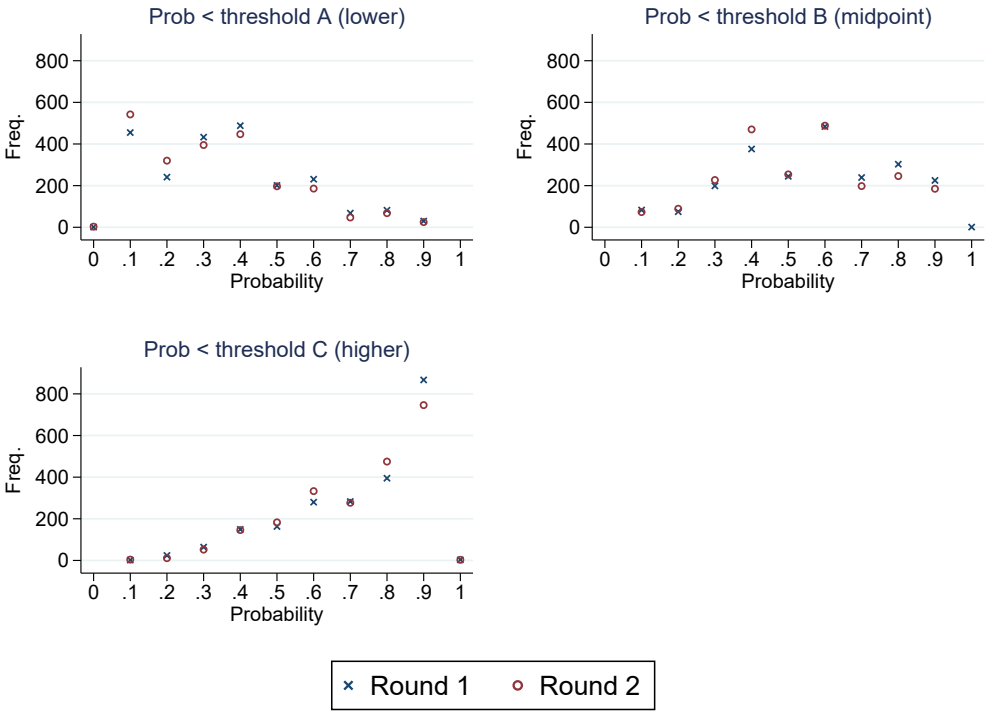
In Table 2, as for the Indian data, we provide descriptive statistics of time-varying characteristics related to income sources and shocks. We divide households into three categories according to the number of household members who report a source of income during the previous month: one, two or three or more members, which account for 38%, 41.3% and 20.7% of the sample, respectively. We also report the proportion of income that comes from regular sources, defined as the share of labor and non-labor income in total household income, excluding occasional labour, monthly CCT subsidies (if any) and transfers. Households with only one working member tend to receive most of their income from regular sources (77.5%), while those with three or more providers are the least likely to do so (34.4%). Income shocks are evenly distributed across these earner categories, similar to the pattern observed in India.

Elicitation of expectations was conducted in a similar fashion to the survey in India; see again Section 2 for a description of the elicitation approach. Figure 5 summarizes the distribution of elicited probabilities by threshold and survey round. Figure 6 displays a high positive correlation between the midpoint of the reported probability distribution and current income in both survey rounds. This relationship is somewhat weaker than in the Indian context, in line with our results on risk and persistence and the fact that expectations here refer to a much shorter time span.

Since the subjective expectations data have not been used before, we provide a detailed analysis and validation in Appendix A.2. Overall, the elicitation of subjective expectations was less precise in the Colombian data. We find a substantially larger degree of logical response errors, although still within reasonable ranges (for instance, around 4% of households report distributions that violate monotonicity). Validation and sample selection leave us with a significantly reduced

[14]The baseline evaluation report can be accessed at https://ifs.org.uk/publications/baseline-report-evaluation-familias-en-accion.

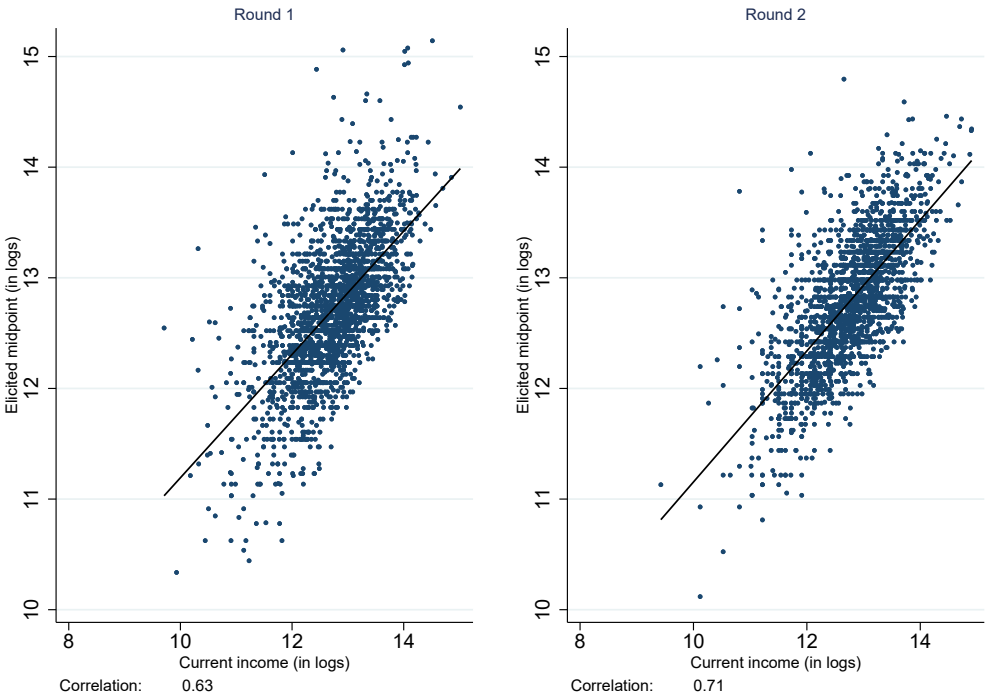
FIGURE 5. Colombia: frequencies by threshold



Note. The figure displays subjective probability frequency plots for each threshold and survey round, as shown in the legend. Probabilities are rounded to the nearest tenth.

balanced panel sample of $N = 2,230$ households. A detailed step-by-step analysis is reported in Appendix A.2. Tables A.5 and A.6 show that these decisions do not imply strong sample selection (at least, based on observable characteristics). Households in the final sample tend to have fewer adults, and household heads are slightly younger (by around a year, on average). They are also slightly less likely to have experienced health or other types of shocks, but are generally very comparable in terms of household composition, income, income sources and education level.

FIGURE 6. Colombia: current income and reported midpoint



Note. The solid black corresponds to the linear regression fit.

5 Results

In this section, we report the results we obtain for both countries when estimating various specifications of the income process. We start with a linear AR(1) process with time effects and consider versions with and without fixed effects, and then augment it with a number of state variables. When these are interacted with current income, persistence is allowed to be heterogeneous in the cross-section. Finally, we consider nonlinear processes of the type introduced in section 3.4. In such models, all the features of the distribution vary across units and over time.

5.1 Linear models

In what follows, we present the estimates of the parameters of the linear process (6), which are obtained from least-squares estimates of the reduced-form parameters in equation (7), using the one-to-one mapping between the two sets of parameters.

5.1.1 India

We begin by reporting estimates using the Indian data in Table 3. The first column contains the estimates of the model without fixed effects, while the second contains the estimates of the model with fixed effects. In addition to the parameters of the model (the persistence parameter ρ, the standard deviation of innovations σ, the residual variance σ_ε^2 and, in the case of the fixed-effect model, the variance of the individual and village-level fixed effects σ_η^2 and $\sigma_{\eta,\text{village}}^2$), for comparability with some of the results we report below, we also include the differences between the 75th and 25th quantiles and between the 90th and 10th quantiles implied by these estimates.

In the model without fixed effects, ρ is close to one and estimated very precisely. The standard deviation of the innovation to the (log) income process is substantive at 0.56, reflected in large values of the interquantile ranges reported. This is a measure of risk, which is identified from the association between the self-reported range of variation of future income and elicited probabilities. The residual variance is estimated at 1.24, which is sizable.

The introduction of fixed effects reduces the degree of persistence from 0.97 to 0.93, which is now significantly different from unity. Fixed effects also play an important role in assessing risk, as the standard error of the income process innovations is reduced from 0.56 to 0.31. The variance of the individual fixed effect at 0.22 (measured as η_i) is one and a half times the variance of the village level fixed effect. These results are surprisingly comparable to those used in standard macro calibrations of the income process based on realized earnings (see, for example, Kaplan and Violante (2010) or Alvarez and Arellano (2022)).

The residual variance is somewhat smaller after fixed effects are introduced, but remains substantial. In particular, it is too large for the residual to be interpreted solely as measurement error in elicited probabilities. This impression is reinforced by the fact that the residual variance in a regression of ℓ_{jit} on r_{jit} with period and unit specific effects is 0.37. Such calculation can be regarded as a lower bound for the measurement error component of the residual in a separable model and implies that a subjective probability of 0.5 would be elicited in the survey with a standard error of one percentage point. The likely presence of additional sources of residual variation provides further motivation for examining the roles of other state variables, nonlinearities, and neglected heterogeneity. However, a variance decomposition with period and unit effects is only suggestive because it preserves the separability between r_{jit} and state variables, while our flexible models emphasize the interactions between the two.

5.1.2 Colombia

In Table 4, we report the results obtained by estimating the same two versions of the linear model (with and without fixed effects) reported in Table 3 for India. In the model without fixed effects, the parameter ρ is estimated to be 0.71. Remarkably, the standard deviation of income innovations is very large at 0.98. It should be remembered that in the Colombian data, future income refers to *next month* rather than *next year*. Annual income is likely to be less volatile than monthly income, although the two economies might be very different. The residual variance is somewhat larger in the Colombian data than in the Indian data, although of comparable magnitude, so the previous comments about the size of the residuals apply here as well.

When adding fixed effects, the estimated ρ is much smaller at 0.5 and the standard deviation of innovations is reduced from 0.98 to 0.65. Moreover, the variance of the individual fixed effect is much larger than in the Indian sample and, similar to India, much larger than the variance of the village component of the fixed effect (four times). Taken together, these results suggest higher risks and lower

	No FE	FE
ρ	0.97	0.93
	(0.94, 1.00)	(0.90, 0.96)
σ	0.56	0.31
	(0.51, 0.60)	(0.29, 0.33)
$IQR_{0.75}$	1.22	0.69
	(1.13, 1.33)	(0.64, 0.74)
$IQR_{0.90}$	2.44	1.38
	(2.25, 2.65)	(1.29, 1.47)
σ^2_η		0.22
		(0.18, 0.27)
σ^2_η village		0.14
		(0.14, 0.19)
σ^2_ε	1.24	1.14
	(1.21, 1.27)	(1.10, 1.18)

Note. The table reports results for the linear model in (7) using the data for India, without fixed effects (and a common intercept) and with fixed effects. We also include year (survey round) dummies in both cases. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 3. India — linear model

persistence in Colombian monthly earnings compared to Indian annual earnings, and greater unobserved heterogeneity in the Colombian data.

	No FE	FE
ρ	0.71	0.50
	(0.67, 0.74)	(0.46, 0.55)
σ	0.98	0.65
	(0.93, 1.03)	(0.63, 0.67)
$IQR_{0.75}$	2.16	1.43
	(2.05, 2.26)	(1.38, 1.48)
$IQR_{0.90}$	4.31	2.86
	(4.10, 4.52)	(2.75, 2.96)
σ^2_η		0.48
		(0.44, 0.52)
σ^2_η village		0.12
		(0.12, 0.17)
σ^2_ε	1.46	1.09
	(1.42, 1.49)	(1.05, 1.12)

Note. The table reports results for the linear model in (7) using the data for Colombia, without fixed effects (and a common intercept) and with fixed effects. We also include year (survey round) and month (interview) dummies in both cases. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 4. Colombia — linear model

5.2 Linear models with additional state variables

We now augment the linear model with time-varying characteristics x_{it}, along the lines of Subsection 3.3 and equation (8). When interacted with current income, we allow for differential subjective persistence along these characteristics. We include sources of income and shocks experienced in the current year as dimensions of x_{it}, as described in Tables 1 and 2 for India and Colombia, respectively. We only report specifications that include fixed effects.

5.2.1 India

The results obtained estimating equation (8) on the India data are reported in Table 5 when introducing indicators of type and number of income sources and in Table 6 when interacting the number of sources with types of shocks.

These results show that persistence is not greatly affected by the presence of additional variables, even when interacted with current income. Households with different sources of income, and households who have in the past year experienced either a health, agricultural or other shock[15] all have subjective persistence around the 0.90 mark. Having said that, however, households with no farm activities have lower levels of persistence.

Remarkably, the introduction of this additional variables does not affect much the variability of the income innovations or of the fixed effects. Similar considerations apply to the residual variance. The conclusion we draw from these tables is that, although marginally significant, the introduction of the interactions with the observable considered, it does not have a large effect on the estimated risk and persistence of the income process.

[15] "Sources" in the India data refer to the number of activities that the household generates income from (farming, agricultural labour, relief work, crafts, trading etc.), health shocks refer to illness or death of a household member, agricultural shocks include crop failure due to disease or floods, and other shocks include events such as job loss or being the victim of crime.

ρ	≤2 sources	3 sources	4+ sources
0% farm	0.87	0.90	0.83
	(0.83, 0.92)	(0.85, 0.95)	(0.76, 0.90)
50% farm	0.91	0.94	0.87
	(0.87, 0.95)	(0.90, 0.98)	(0.80, 0.93)
75% farm	0.93	0.96	0.89
	(0.88, 0.98)	(0.92, 1.01)	(0.82, 0.96)
σ		0.30	
		(0.28, 0.32)	
$IQR_{0.90}$		1.33	
		(1.24, 1.41)	
σ^2_η		0.23	
		(0.19, 0.28)	
σ^2_η village		0.13	
		(0.13, 0.18)	
σ^2_ε		1.12	
		(1.07, 1.16)	

Note. The table reports results for India for the linear model (7) in augmented with household-level characteristics, along the lines of (8). "Farm" refers to the proportion of current income obtained from farming-related activities; see Table 1 for a full description of these variables. We also include year (survey round) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 5. India — linear model augmented with household characteristics (% of income from farming)

ρ	≤2 sources	3 sources	4+ sources
No shock	0.87	0.91	0.83
	(0.79, 0.95)	(0.84, 0.99)	(0.74, 0.93)
Health	0.92	0.97	0.89
	(0.86, 0.98)	(0.91, 1.04)	(0.81, 0.97)
Agricultural	0.90	0.97	0.87
	(0.86, 0.95)	(0.92, 1.01)	(0.81, 0.94)
Other	0.99	1.04	0.97
	(0.88, 1.09)	(0.93, 1.14)	(0.84, 1.08)
σ		0.30	
		(0.28, 0.32)	
$IQR_{0.90}$		1.34	
		(1.23, 1.42)	
σ^2_η		0.25	
		(0.22, 0.32)	
σ^2_η village		0.15	
		(0.15, 0.21)	
σ^2_ε		1.13	
		(1.08, 1.16)	

Note. The table reports results for India for the linear model in (7) augmented with household-level characteristics, along the lines of (8). See Table 1 for a full description of these variables. We also include year (survey round) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 6. India — linear model augmented with household characteristics (shocks and income sources)

5.2.2 Colombia

In the Colombian data, we note more substantial differences in persistence according to the number of sources of income than in India. We observe that for three or more working members there is higher persistence. This seems to be a case of income diversification, which makes a lot of sense for Colombia given that predictions are one-month ahead. This is particularly true for households with a regular (non-occasional) stream of income, as can be seen in Table 7.

It is noteworthy that the introduction of the additional controls interacted with income does not make much difference to the size of the uncertainty, which remains more or less at the same level (0.64) and to the variability of both individual and village level fixed effects. The stability of these coefficients and the limited variability of the persistence estimates are an indication of the fact that these observables play a limited role in this model.

The most noticeable difference between these results and those obtained for India is the size of the persistence coefficient. In the case of Colombia, although some variability is observed by income sources, persistence is never larger than 0.65, while in India is never below 0.83. We also notice that the idiosyncratic variability of fixed effects is considerably larger in Colombia, being almost twice as large as in India. Instead, the variability of village level fixed effects is roughly similar (around 0.12). An implication of this finding is that the variability across villages accounts for a larger fraction of the fixed effects variability in India.

ρ	1 earner	2 earners	3+ earners
0% regular	0.34	0.36	0.48
	(0.21, 0.48)	(0.24, 0.47)	(0.34, 0.63)
75% regular	0.51	0.52	0.61
	(0.43, 0.58)	(0.43, 0.59)	(0.51, 0.71)
100% regular	0.56	0.57	0.65
	(0.49, 0.63)	(0.48, 0.66)	(0.55, 0.76)
σ		0.64	
		(0.62, 0.67)	
$IQR_{0.90}$		2.83	
		(2.72, 2.92)	
σ^2_η		0.48	
		(0.45, 0.52)	
σ^2_η village		0.11	
		(0.12, 0.16)	
σ^2_ε		1.08	
		(1.04, 1.11)	

Note. The table reports results for Colombia for the linear model in (7) augmented with household-level characteristics, along the lines of (8). See Table 2 for a full description of these variables. We also include year (survey round) and month (interview) dummies. The R^2 are adjusted for the presence of fixed effects. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 7. Colombia — linear model augmented with household characteristics (proportion of regular income and income sources)

ρ	1 earner	2 earners	3+ earners
No shock	0.54	0.55	0.58
	(0.47, 0.62)	(0.47, 0.63)	(0.48, 0.68)
Health	0.64	0.65	0.67
	(0.50, 0.77)	(0.51, 0.79)	(0.53, 0.81)
Other	0.44	0.46	0.48
	(0.32, 0.56)	(0.33, 0.58)	(0.34, 0.61)
σ		0.65	
		(0.62, 0.67)	
$IQR_{0.90}$		2.84	
		(2.73, 2.93)	
σ^2_η		0.48	
		(0.45, 0.53)	
σ^2_η village		0.11	
		(0.12, 0.16)	
σ^2_ε		1.09	
		(1.04, 1.11)	

Note. The table reports results for Colombia for the linear model (7) in augmented with household-level characteristics, along the lines of (8). See Table 2 for a full description of these variables. We also include year (survey round) and month (interview) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 8. Colombia — linear model augmented with household characteristics (income shocks and income sources)

5.3 Nonlinear models

We now turn to the central empirical results of the paper. In this section, we discuss the estimation of the nonlinear model with additive fixed effects:

$$\ell_{jit} = \beta_0^\dagger(s_{jit}) + \beta_1^\dagger(s_{jit})\psi(y_{it}) + \eta_i + \varepsilon_{jit}, \tag{20}$$

which corresponds to equation (16) with $\beta_2^\dagger(\cdot) = 1$. We report results for the most parsimonious specification that would still allow for nonlinear persistence and skewness — that is, we choose $L = 3$ for $\beta_0^\dagger(\cdot)$ and $\beta_1^\dagger(\cdot)$ (cubic splines with two boundary knots and one intermediate knot) and $\psi(\cdot)$ of order 1 (current income enters in log levels, but since $s_{jit} = r_{jit} - y_{i,t}$ quadratic income terms are also involved). This configuration is in what follows referred to as the *baseline* specification for the nonlinear model.

We experimented with higher values of L and higher order polynomials. While for $L > 3$ we obtained qualitatively similar results, as long as the position of the knots was judiciously chosen, the use of higher-order $\psi(\cdot)$ terms required substantial penalization to avoid unstable results. We also experimented with nonlinear models involving additional state variables, but since the interaction terms did not contribute much, in line with what we saw for linear models, we do not present them here.

In our data, identification of nonlinear persistence comes directly from the association between current income and the shape of the distribution of subjective probabilities of future income, net of fixed effects. In particular, to obtain the results we present, it is key to consider distributional models that are flexible enough to allow for conditional skewness that may change with current income and unobserved heterogeneity.

Overall, the linear autoregressive model is soundly rejected on both the Indian and Colombian data. Moreover, similar patterns of heterogeneous risks and nonlinear persistence emerge in the two data sets. This is particularly remarkable in view of the differences between the two surveys (annual vs monthly) and the characteristics of their underlying populations.

5.3.1 India

Table 9 presents the results obtained in estimating the nonlinear model with additive effects on the Indian data. Firstly, with regard to risk, it is noticeable that *dispersion risk* decreases with current income (interquantile range measures for the 90th income percentile are two-thirds of those for the 10th percentile), while *skewness risk* increases moderately. That is, the rich have less dispersion risk but more skewness risk than the poor.

Turning to persistence, we observe the presence of nonlinear persistence, which depends on both the percentile of current income and the rank of the quantile shock to next-period's income. Persistence is close to one for high-income households throughout, but only when hit by a bad shock for low-income households. When a good shock hits a low-income household, persistence is much lower. This pattern, which is depicted in Figure 7, features prominently in our results and is only partially consistent with the nonlinear persistence reported in Arellano et al. (2017), who found reduced persistence not only at the bottom of the income distribution (with good shocks) but also at the top of the income distribution (with bad shocks). Those differences do not necessarily imply a contradiction between results based on subjective expectations and those based on realized incomes because the populations of reference in the two studies are very different. In our developing country databases all households are poor by comparison to PSID households.

An economic implication of the nonlinear income process comes from the fact that a positive shock for lower-income households reduces the persistence of the past and is therefore beneficial for those households in terms of expected future income. Relative to the predictions of a linear income process, this asymmetry will induce lower saving and higher consumption at younger ages for self-insured low-income households. However, for higher-income households, given the estimated process, the opposite effect (associated to negative shocks) would not be expected to happen.

The nonlinear persistence that we find among the poorest households is consistent with a poverty trap interpretation. When income is too low it is difficult to escape poverty, but a large positive shock can weaken the weight of the past

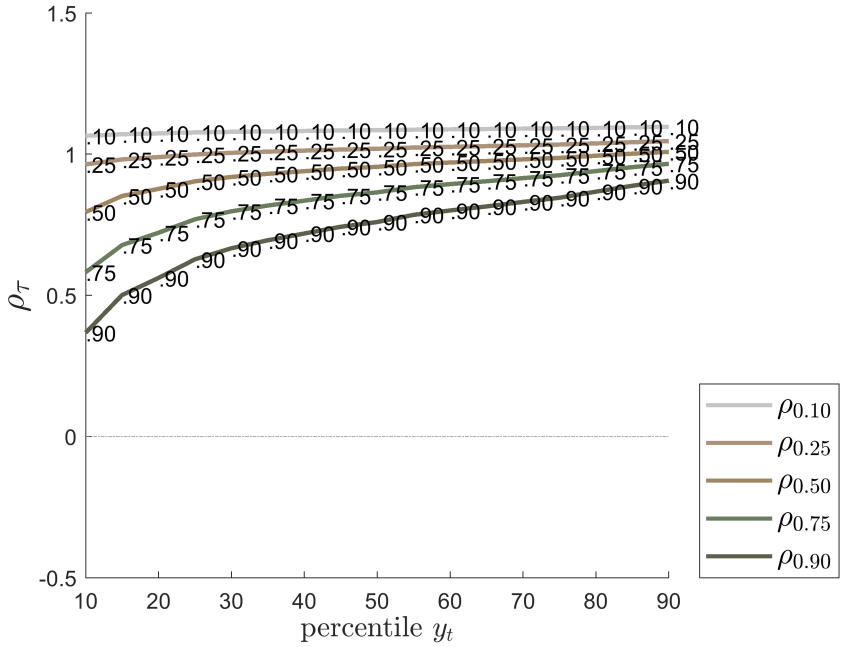
	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	0.79	0.60	0.52
	(0.72, 0.90)	(0.54, 0.63)	(0.45, 0.55)
$IQR_{0.90}$	1.61	1.23	1.05
	(1.48, 1.85)	(1.15, 1.31)	(0.93, 1.14)
$SK_{0.90}$	−0.02	−0.10	−0.14
	(−0.15, 0.06)	(−0.20, −0.04)	(−0.28, −0.04)
$\rho_{\tau 0.25}$	0.96	1.02	1.05
	(0.92, 1.05)	(0.99, 1.06)	(1.00, 1.08)
$\rho_{\tau 0.50}$	0.79	0.96	1.01
	(0.72, 0.86)	(0.92, 0.98)	(0.97, 1.03)
$\rho_{\tau 0.75}$	0.58	0.86	0.97
	(0.38, 0.71)	(0.82, 0.89)	(0.91, 0.99)
σ^2_η		0.23	
		(0.19, 0.28)	
σ^2_η village		0.14	
		(0.13, 0.19)	
σ^2_ε		1.11	
		(1.06, 1.14)	

Note. The table reports results for India for the flexible model with additive fixed effects in (20). We also include year (survey round) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 9. India — flexible model (additive fixed effects)

history and get the household (persistently) off the hook at a higher income level.[16] We find it quite interesting that this kind of poverty trap dynamics seems to be reflected in the subjective income expectations of poor households.

[16]See Banerjee, Duflo, Goldberg, Karlan, Osei, Parienté, Shapiro, Thuysbaert, and Udry (2015) for evidence on how a multifaceted program can help the extreme poor to persistently increase their income; and Genicot and Ray (2017) for an aspirations-based theory of poverty traps and references to the earlier theoretical literature.



Note. The figure reports estimates of nonlinear persistence for India for the flexible model with additive fixed effects in (20). Specifications also include year (survey round) dummies. See Figure C.3 for pointwise confidence bands.

FIGURE 7. India — flexible model, nonlinear persistence (additive fixed effects)

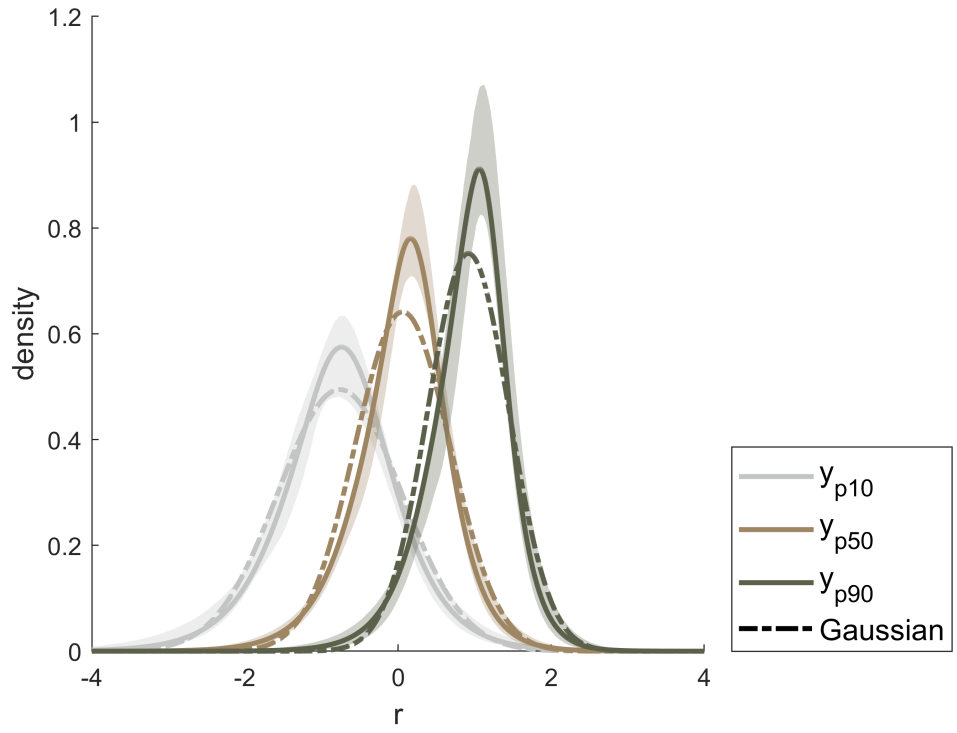
All these summary measures are computed for the model's probability distributions evaluated at the median value of the fixed effects. We extend the analysis to other percentiles (corresponding to a normal distribution with the estimated variance of the fixed effects) in Figure C.1 (see Appendix C) and obtain a similar pattern of nonlinear persistence together with an additional pattern of unobserved heterogeneity. Specifically, as the selected percentile of the effects increases, overall persistence increases and the amount of persistence for different shocks becomes more compressed, with a flatter gradient along current income for large shocks.

Probability density functions for India. Figure 8 shows estimated conditional probability density functions (*pdf*s) at the 10th, 50th and 90th percentiles of current income for the baseline model. They are calculated by numerical differentiation

of the corresponding estimated cumulative probabilities. Estimated *cdf*s with or without rearrangement coincide since there are no instances of non-monotonicities in the baseline specification. Figure 8 also shows block-bootstrap point-wise confidence bands and Normal *pdf*s with the same empirical mean and variance for comparisons.

As expected, the predictive subjective density for poorer households is shifted to the left relative to that of richer households, indicating that at any given reference level for future income, poorer households tend to assign a higher probability to their future income falling below that level. Moreover, consistent with the results in Table 9, subjective predictive densities tend to be more symmetric and are noticeably more dispersed for the poor. Beyond non-normality, the figure also portrays more pronounced differences between the current rich and the current poor in terms of relatively bad and relatively good outcomes. These differences are suggestive of nonlinear persistence,[17] which is more relevant for the currently poor, consistently with the patterns we find in Table 9.

[17] Recall that, according to the chain rule in equation (15), nonlinear persistence can be obtained as the derivative of the *cdf* with respect to y relative to the derivative of the *cdf* with respect to r.



Note. The figure shows estimated *pdf*s at the 10th, 50th and 90th percentiles of current income for India, calculated by numerical differentiation on the estimated (conditional) cumulative probabilities. Shared areas are 90% pointwise confidence bands using block bootstrap (1000 repetitions). Dotted lines correspond to Normal *pdf*s with the same mean and variance as the empirical *pdf*s of the same color. The range of variation is standardized (future) log income, see Appendix B.4.

FIGURE 8. India — flexible model, probability density function (additive fixed effects)

5.3.2 Colombia

Table 10 presents the results for the nonlinear model with additive effects on the Colombian data. Similar to India, we observe dispersion and skewness decreasing with current income (decreasing dispersion risk and increasing skewness risk). However, while in India we found no skewness at the bottom of the income distribution and negative skewness at the top, in Colombia we find positive skewness at the bottom and no skewness at the top.

Regarding persistence, although at lower levels than in India (similar to the linear model), we find the same pattern of nonlinearities, with persistence decreasing

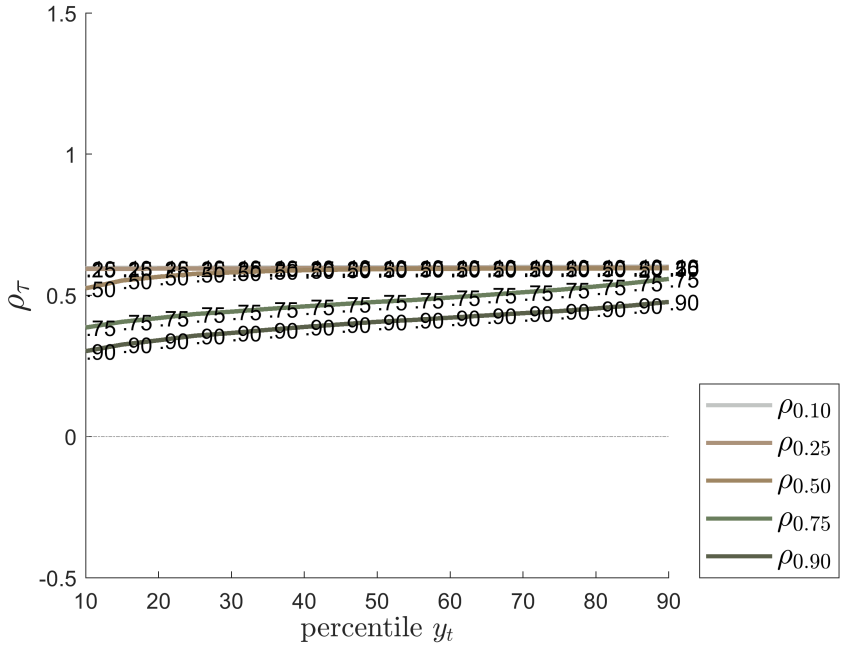
	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	1.48	1.34	1.28
	(1.39, 1.58)	(1.26, 1.40)	(1.21, 1.36)
$IQR_{0.90}$	2.97	2.75	2.63
	(2.80, 3.13)	(2.63, 2.86)	(2.50, 2.77)
$SK_{0.90}$	0.13	0.08	0.04
	(0.05, 0.19)	(0.03, 0.12)	(−0.01, 0.08)
$\rho_{\tau 0.25}$	0.59	0.60	0.60
	(0.54, 0.65)	(0.53, 0.67)	(0.48, 0.69)
$\rho_{\tau 0.50}$	0.52	0.59	0.60
	(0.42, 0.61)	(0.54, 0.64)	(0.52, 0.66)
$\rho_{\tau 0.75}$	0.39	0.48	0.56
	(0.30, 0.47)	(0.41, 0.54)	(0.50, 0.62)
σ^2_η		0.47	
		(0.44, 0.52)	
σ^2_η village		0.12	
		(0.12, 0.17)	
σ^2_ε		1.09	
		(1.05, 1.12)	

Note. The table reports results for Colombia for the flexible model with additive fixed effects in (20). We also include year (survey round) and month (interview) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 10. Colombia — flexible model (additive fixed effects)

with relatively good shocks for low income households but not for high income households (Table 10 and Figure 9). The impact of unobserved heterogeneity on persistence is also similar to the one for India; see Figure C.2 in Appendix C.

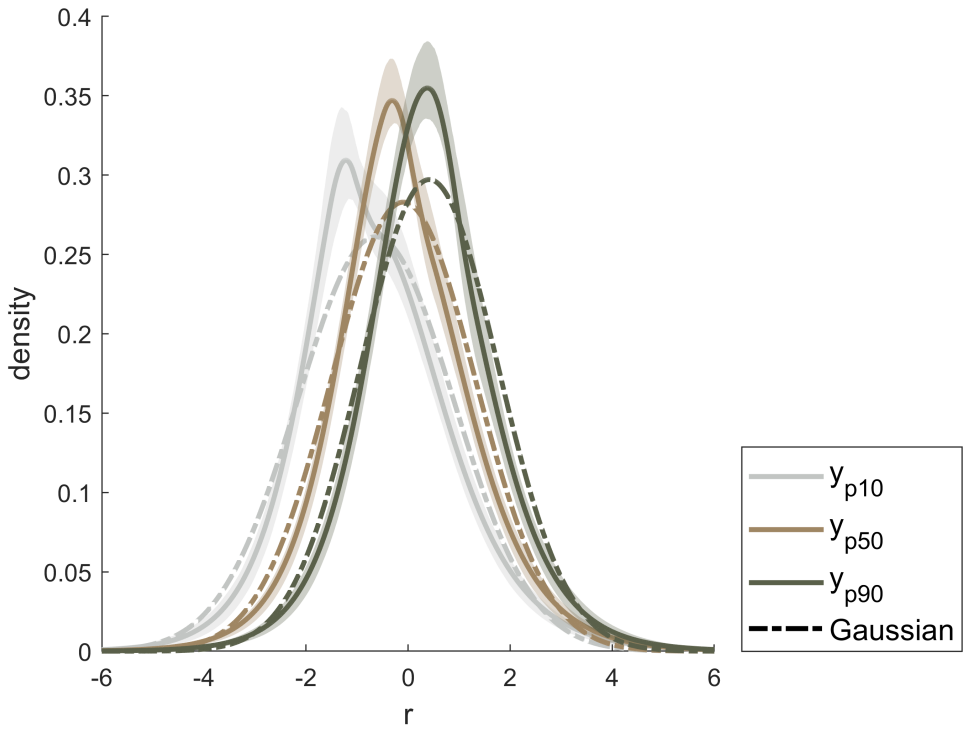
Probability density functions for Colombia. We also observe marked departures from normality in Colombia according to Figure 10, which depicts estimated *pdf*s together with Normal distribution fits. The estimated densities are consistent



Note. The figure reports estimates of nonlinear persistence for Colombia for the flexible model with additive fixed effects in (20). Specifications also include year (survey round) dummies. See Figure C.4 for pointwise confidence bands.

FIGURE 9. Colombia — flexible model, nonlinear persistence (additive fixed effects)

with the pattern in Table 10 of decreasing dispersion risk as we move along the income gradient and feature prominent deviations from normality, more so for poorer households.



Note. The figure shows estimated *pdf*s at the 10th, 50th and 90th percentiles of current income for Colombia, calculated by numerical differentiation on the estimated (conditional) cumulative probabilities. Shared areas are 90% pointwise confidence bands using block bootstrap (1000 repetitions). Dotted lines correspond to Normal *pdf*s with the same mean and variance as the empirical *pdfs* of the same color. The range of variation is standardized (future) log income, see Appendix B.4.

FIGURE 10. Colombia — flexible model, probability density function (additive fixed effects)

5.4 Generalizing heterogeneity patterns: interacted fixed effects

In this section, we discuss the estimation of the nonlinear model (16) with interacted fixed effects, in which $\beta_2^\dagger(s_{jit})$ is allowed to depend on s_{jit}. In these models log odd ratios can vary differentially with fixed effects and therefore allow for a greater distributional role of unobserved heterogeneity in accounting for nonlinearities.

In line with the nonlinear estimates with additive effects, we report results for a parsimonious specification where we choose $L = 3$ for $\beta_0^\dagger(\cdot)$ and $\beta_1^\dagger(\cdot)$, $L = 2$ for $\beta_2^\dagger(\cdot)$ and $\psi(\cdot)$ of order 1. Thus, the model contains a total of 6 parameters — two

in the intercept function, three in the interactions with current income, and one in the multiplicative term interacted with the fixed effect. We resort to the linear TSLS estimator introduced in Section 3.5.2 (which does not impose the restrictions in equation (19)) using a full set of first-stage interaction terms.[18]

Tables 11 and 12 report the results for India and Colombia, respectively. Starting with the results for India, we observe some noticeable differences relative to the nonlinear additive model estimates in Table 9. First, dispersion risk is now smaller overall, although it is still decreasing with current income. Thus, it appears that a larger fraction of the spread in the subjective probability distributions is now accounted for by unobserved heterogeneity as opposed to risk. Secondly, negative skewness is now more prominent overall, while the increase in skewness risk with current income is much reduced. Finally, although the pattern of nonlinear persistence remains the same, there is a smaller reduction in persistence at the bottom of the income distribution in the presence of a positive shock. The results for Colombia in Table 12 tell a similar story relative to those in Table 10 for the nonlinear additive model.

State dependence versus unobserved heterogeneity. We have found that state dependence and unobserved heterogeneity compete as sources to explain persistence, not only in linear models (the comparison between columns 1 and 2 in tables 3 and 4), but also in the case of nonlinear persistence when fixed effects are allowed a flexible distributional role.[19] Our results show that both state dependence and unobserved heterogeneity matter, linearly and nonlinearly, and illustrate how to

[18]Remember that we also include year indicators in all specifications, and that the TSLS estimator on the transformed equation (19) requires us to account for additional "included" regressors (even if the nonlinear restrictions are not imposed). In the case of Colombia, relative to the nonlinear estimates with additive fixed effects, we excluded survey (month) indicators, which would increase the regressor set by 16 additional coefficients and tend to introduce instability in the estimates.

[19]Using a different model for realized outcomes, Almuzara (2020) considers a related problem of distinguishing between nonlinear state dependence and (variance) unobserved heterogeneity. He shows that a fixed effect in the variance of transitory shocks may give rise to spurious nonlinear dynamics.

	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	0.56	0.46	0.42
	(0.49, 0.79)	(0.39, 0.56)	(0.33, 0.48)
$IQR_{0.90}$	1.31	1.04	0.90
	(1.04, 3.32)	(0.83, 1.50)	(0.70, 1.12)
$SK_{0.90}$	−0.25	−0.29	−0.29
	(−0.70, −0.04)	(−0.50, −0.11)	(−0.45, −0.12)
$\rho_{\tau 0.25}$	1.00	1.05	1.07
	(0.93, 1.11)	(1.01, 1.10)	(1.03, 1.10)
$\rho_{\tau 0.50}$	0.93	1.01	1.04
	(0.83, 0.97)	(0.95, 1.03)	(0.99, 1.06)
$\rho_{\tau 0.75}$	0.82	0.97	1.02
	(0.63, 0.88)	(0.89, 0.99)	(0.95, 1.04)
σ^2_η		0.49	
		(0.38, 0.63)	
σ^2_η village		0.19	
		(0.18, 0.29)	
σ^2_ε		1.10	
		(1.01, 1.26)	

Note. The table reports results for India for the flexible model in (16). We also include year (survey round) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 11. India — flexible model (multiplicative fixed effects)

quantify the relative contributions of each one to different features of a distributional income process estimated from subjective expectations data.

	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	1.91	1.62	1.52
	(1.22, 4.19)	(1.11, 3.00)	(1.10, 2.41)
$IQR_{0.90}$	3.85	3.57	3.48
	(2.49, 8.13)	(2.39, 6.74)	(2.37, 6.18)
$SK_{0.90}$	0.37	0.27	0.16
	(0.21, 0.56)	(0.13, 0.50)	(0.05, 0.37)
$\rho_{\tau 0.25}$	0.59	0.49	0.38
	(0.46, 0.69)	(0.26, 0.65)	(−0.07, 0.62)
$\rho_{\tau 0.50}$	0.50	0.58	0.49
	(−0.31, 0.68)	(0.41, 0.68)	(0.20, 0.65)
$\rho_{\tau 0.75}$	0.19	0.26	0.39
	(−1.24, 0.53)	(−0.72, 0.57)	(−0.39, 0.63)
σ^2_η		0.47	
		(0.41, 0.58)	
σ^2_η village		0.11	
		(0.11, 0.17)	
σ^2_ε		1.10	
		(1.05, 1.23)	

Note. The table reports results for Colombia for the flexible model in (16). We also include year (survey round) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE 12. Colombia — flexible model (multiplicative fixed effects)

6 Conclusion

We have developed an econometric framework for modeling income risk and heterogeneity from the responses to subjective expectation questions of Indian and Colombian households. A main conclusion is that linear income processes are soundly rejected in both datasets. Subjective income distributions feature heteroskedasticity, conditional skewness, and nonlinear persistence. We find a negative association between conditional dispersion and current income, and between conditional skewness and current income. We also find that persistence diminishes for poor households experiencing large positive shocks, but not for richer households experiencing large negative shocks.

Unobserved heterogeneity matters and is composed of both household specific and aggregate-level factors. We find that state dependence and unobserved heterogeneity compete as explanations of risk and persistence, both linearly and nonlinearly, which emphasizes the importance of allowing for flexible distributional unobserved heterogeneity to be able to capture their relative contributions. We also explored to what extent not only current income but also its sources matter for risk, thereby calling for a larger state space than is common in the literature, and found only moderate evidence for the role of those additional state variables.

Taken together, our results suggest complex and heterogeneous patterns of transmission of income shocks to consumption, involving precautionary dispersion and skewness motives, which depend on the household position in the income distribution.

References

Almuzara, M. (2020): "Heterogeneity in Transitory Income Risk." Working paper.

Altonji, J.G., H. D. and I. Vidangos (2023): "Individual earnings and family income: Dynamics and distribution." *Review of Economic Dynamics*, 49, 225–250.

Alvarez, J. and M. Arellano (2022): "Robust likelihood estimation of dynamic panel data models." *Journal of Econometrics*, 226, 21–61.

Arellano, M. (2014): "Uncertainty, Persistence, and Heterogeneity: A Panel Data Perspective." *Journal of the European Economic Association*, 12, 1127–1153.

Arellano, M., R. Blundell, and S. Bonhomme (2017): "Earnings and consumption dynamics: A nonlinear panel data framework." *Econometrica*, 95, 693–734.

Arellano, M. and S. Bonhomme (2012): "Identifying distributional characteristics in random coefficients panel data models." *Review of Economic Studies*, 79, 987–1020.

——— (2016): "Nonlinear panel data estimation via quantile regressions." *Econometrics Journal*, 19, C61–C94.

Arellano, M., S. Bonhomme, M. De Vera, L. Hospido, and S. Wei (2022): "Income Risk Inequality: Evidence from Spanish Administrative Records." *Quantitative Economics*, 13, 1747–1801.

Attanasio, O. (2009): "Expectations and Perceptions in Developing Countries: Their Measurement and Their Use." *American Economic Review: Papers and Proceedings*, 99, 87–92.

Attanasio, O. and B. Augsburg (2016): "Subjective Expectations and Income Processes in Rural India," *Economica*, 83, 416–442.

Attanasio, O., E. Battistin, and A. Mesnard (2012): "Food and Cash Transfers: Evidence from Colombia." *Economic Journal*, 122, 92–124.

Augsburg, B. (2009): *Microfinance - Greater Good or Lesser Evil?*, PhD thesis, Maastricht Graduate School of Governance, University of Maastricht.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry (2015): "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." *Science*, 348.

Chamberlain, G. (1992): "Efficiency bounds for semiparametric regression." *Econometrica*, 60, 567–596.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models." in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, vol. 6B, chap. 76.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantile and probability curves without crossing." *Econometrica*, 78, 1093–1125.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on Counterfactual Distributions." *Econometrica*, 81, 2205–2268.

Cox, D. R. and E. J. Snell (1970): *Analysis of Binary Data*, Cambridge University Press, second ed.

De Nardi, M., F. G. and G. Paz-Pardo (2020): "Nonlinear Household Earnings Dynamics, Self-Insurance, and Welfare." *Journal of the European Economic Association*, 18, 890–926.

Delavande, A. (2023): "Expectations in development economics." in *Handbook of Economic Expectations*, ed. by T. G. Bachmann, R. and W. van der Klaauw, 261–291.

Delavande, A., X. Giné, and D. McKenzie (2011): "Measuring subjective expectations in developing countries: A critical review and new evidence." *Journal of Development Economics*, 94, 151–163.

Delavande, A. and S. Rohwedder (2008): "Eliciting Subjective Probabilities in Internet Surveys." *Public Opinion Quarterly*, 72, 866–891.

Dominitz, J. (1998): "Earnings Expectations, Revisions, and Realizations." *Review of Economics and Statistics*, 80, 374–388.

——— (2001): "Estimation of income expectations models using expectations and realization data." *Journal of Econometrics*, 102, 165–195.

Dominitz, J. and C. F. Manski (1996): "Eliciting Student Expectations of the Returns to Schooling." *Journal of Human Resources*, 31, 1–26.

——— (1997a): "Perceptions of economic insecurity: evidence from the survey of economic expectations," *Public Opinion Quarterly*, 61, 261–287.

——— (1997b): "Using expectations data to study subjective income expectations." *Journal of the American Statistical Association*, 92, 855–867.

Evdokimov, K. (2010): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity." Working paper.

Foresi, S. and F. Peracchi (1995): "The conditional distribution of excess returns: An empirical analysis." *Journal of the American Statistical Association*, 90, 451–466.

Genicot, G. and D. Ray (2017): "Aspirations and Inequality." *Econometrica*, 85, 489–519.

Golosov, M. and A. Tsyvinski (2015): "Policy Implications of Dynamic Public Finance." *Annual Review of Economics*, 7, 147–171.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song (2021): "What Do Data on Millions of U.S. Workers Say About Labor Income Risk?" *Econometrica*, 89, 2303–2339.

Hall, R. and F. Mishkin (1982): "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households." *Econometrica*, 50, 461–481.

Hu, Y. (2017): "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics." *Journal of Econometrics*, 200, 154–168.

Kaplan, G. and G. L. Violante (2010): "How much consumption insurance beyond self-insurance?" *American Economic Journal: Macroeconomics*, 2, 53–87.

Manski, C. (2004): "Measuring Expectations." *Econometrica*, 72, 1329–1376.

——— (2018): "Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise." in *NBER Macroeconomics Annual*, ed. by M. Eichenbaum and J. Parker, University of Chicago Press.

Meghir, C. and L. Pistaferri (2011): "Earnings, Consumption, and Life Cycle Choices," in *Handbook of Labor Economics*, ed. by D. Card and O. Ashenfelter, Elsevier, vol. 4B, 773–854.

Morgan, M. G. and M. Henrion (1990): *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press.

Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects." *Econometrica*, 49, 1417–1426.

Schennach, S. M. (2022): "Measurement Systems." *Journal of Economic Literature*, 60, 1223–1263.

APPENDIX

A Data

A.1 India – sample selection and validation

This section describes sample selection and validation of the subjective expectations data in detail. For the most part, it mirrors the analysis in Section 1 in Attanasio and Augsburg (2016). However, we update some of the criteria used and review additional sample selection decisions needed.

Tables A.1 and A.2 can be understood as an extended version of Tables 2 and 3 (respectively) reported in Attanasio and Augsburg (2016).

Table A.1 details basic sample selection steps and resulting sample sizes. For instance, we exclude observations with missing income or at least one reported probability. We also report the number of households for which either the elicited subjective lower or upper bound on future income is missing, although we do not exclude these from the final sample.[20] We also exclude from the sample some extreme reports of current household income under the category "implausible income", which are likely to correspond to survey measurement error.[21]

The total number of unique households drops to 930 and 877 in the first and second rounds, respectively. Relative to these, we study bunching of the reported probabilities at the 0%, 50% and 100% marks in Table A.2, and note substantial bunching at 50% for the midpoint (especially in the first round).

Keeping only households present in both rounds implies a balanced panel with 789 unique households. We further drop households who report at least one

[20] A total of 1,041 households were originally interviewed in the first round. We drop five observations who were asked about monthly rather than yearly income. In the remaining rows, minor differences with respect to those reported in Attanasio and Augsburg (2016) are due to slightly different/updated criteria.

[21] In particular, these correspond to reports outside the range $[0.5 \times rmin, 2 \times rmax]$, where *rmin* and *rmax* are the reported lower and upper bound on subjective income distributions (respectively), and are replaced by r_A and r_C, respectively, if missing. Unlike in the Colombian data, as reported in Appendix A.2, using looser or more strict "cutoffs" does not lead to substantial changes in sample sizes.

probability as equal to zero or one or whose elicited subjective *cdf* is not *strictly* monotonic. This further reduces the final sample size to $N = 770$ households. We present robustness checks keeping those households in Appendix D.1.

TABLE A.1. India: response rates and sample sizes

	Round 1	Round 2
Total number of observations	1036	947
Missing income	2	1
Missing either Min or Max	11	11
Missing at least one probability	22	11
Wrong — direction	3	5
Wrong — violation of monotonicity	9	19
Implausible thresholds	63	2
Implausible income	22	41
Available observations	926	873
Balanced panel (robustness)	789	
At least one probability is 0 or 1	9	11
At least two prob. are equal	3	2
Balanced panel (final)	770	

Note. "Wrong — direction" refers to households that in a given survey report $\Pr(y_{t+1} \geq r_C) > \Pr(y_{t+1} \geq r_B) > \Pr(y_{t+1} \geq r_A)$, among those with no missing probabilities. "Wrong — violation of monotonicity" refers to weak monotonicity violations. "Implausible thresholds" refers to households for which r_A, r_B or r_C is missing, among those who report no missing probabilities, households for which $r_B < r_A$ or $r_C < r_B$, and households with implausibly large interval differences between $r_B - r_A$ and $r_C - r_B$. "Implausible income" refers to households that report income outside $[0.5\times rmin, 2\times rmax]$, where *rmin* and *rmax* are replaced by r_A and r_C, respectively, if missing.

TABLE A.2. India: bunching

	Round 1	Round 2
Threshold A (Lower)		
0%	6	10
50%	1	20
100%	2	0
Threshold B (Midpoint)		
0%	0	0
50%	503	142
100%	2	0
Threshold C (Higher)		
0%	0	0
50%	44	102
100%	3	1
Available observations	926	873

Note. The table shows the number of respondents who reported 0%, 50% and 100% probabilities in each survey round.

A.2 Colombia – sample selection and validation

We repeat the same analysis as for the Indian data (Appendix A.1) here. Tables A.3 and A.4 summarize sample selection decisions and validation of the expectations data and bunching, respectively.

As shown in Table A.3, of the original sample of 11,462 households interviewed during the 2002 baseline survey, 10,743 were re-interviewed in 2003 and 9,463 in 2005/06, of which 9,221 provided information on household income during the first survey round and 7,517 during the second. The fraction of households with missing reported probabilities or extreme values household income is also larger than in India.[22] Relative to the analysis for India, here we include an additional step

[22]In the Colombian data, using looser or more strict rules for "implausible income" (described in Table A.3) leads to substantially larger and smaller sample sizes, respectively. For instance, using an

on "missing covariates", where we exclude a few households for which covariates used in the main analysis (village, number of income sources and the proportion of income from farming sources) were missing.

The remaining rows in Table A.3 and Table A.4 correspond to the validation exercise on the subjective expectations data, similar to the one performed for India in Appendix A.1 and in Attanasio and Augsburg (2016). The final balanced sample we use in the main analysis has $N = 2,230$ unique households, after excluding those who report probabilities equal to zero or one or answers that violate strict monotonicity of (subjective) *cdf*s. We report our main results keeping these observations in Appendix D.2. Finally, Tables A.5 and A.6 report differences in observable characteristics between households in the final dataset and those available but excluded at after validation.

Since the subjective expectations data in Colombia has not been used before, we now elaborate on validation:

- *Logical response errors*. Table A.3 shows that in the first survey round 350 households provided answers that violated monotonicity and 37 provided responses that adhered to monotonicity but were "inverted", in that probabilities were non-decreasing, rather than non-increasing.[23] These figures imply violations of around 4% of those that gave responses to the probabilities, somewhat higher than in the Indian context where the logical response error was around 1%, but comparable to other studies. Dominitz and Manski (1997a), for example, report violations for almost 5% of their sample, and almost twice that number when including respondents where prompting happened (in that their responses would have initially been classified as logical response errors, but changed responses after having been prompted; such prompting was not allowed in either of the contexts considered here).

interval given by $[0.2 \times rmin, 5 \times rmax]$ allows us to keep approximately around 200 additional unique households. Naturally, repeating the analysis with this rather permissive rule tends to exaggerate the features (nonlinear persistence, skewness, etc) we study.

[23] Recall that, as described in Figure 1, households were asked to report probabilities of the form $\Pr(y_{t+1} \geq r)$.

- *Bunching of percentages*. Table A.4 reports on the extent to which households bunch at the 0%, 50%, or 100% probability marks for the different thresholds. In the first round, there is substantial bunching at 0% for the lowest and 100% for the highest thresholds (around 14% of responses, compared to a negligible amount in India), which suggests some households might not have understood the prompts correctly or that the elicited expected income range is not accurate. There is also apparent bunching at 50% for the midpoint, a common feature with subjective probability data also present in the Indian data. There is a more muted presence of these issues in the second wave data.

Even though these data display a higher degree of logical response errors and bunching than in India, we conclude that the responses provided in the Colombian data conform for the most part to the basic probability laws, and seem to suggest substantial coherency and variability, in that most respondents appear to have understood the instructions and provided thoughtful responses.

TABLE A.3. Colombia: response rates and sample sizes

	Round 1	Round 2
Total number of observations	10743	9463
Missing income	1522	1946
Missing either Min or Max	1294	958
Missing at least one probability	1361	964
Wrong — direction	37	33
Wrong — violation of monotonicity	350	291
Implausible thresholds	24	21
Implausible income	633	390
Available observations	7262	6295
Missing covariates	38	25
Balanced panel (robustness)	4420	
At least one probability is 0 or 1	2005	1434
At least two elicited probabilities are equal	866	600
Balanced panel (final)	2230	

Note. "Wrong — direction" refers to households that in a given survey report $\Pr(y_{t+1} \geq r_C) > \Pr(y_{t+1} \geq r_B) > \Pr(y_{t+1} \geq r_A)$, among those with no missing probabilities. "Wrong — violation of monotonicity" refers to weak monotonicity violations. "Implausible thresholds" refers to households for which r_A, r_B or r_C is missing, among those who report no missing probabilities, households for which $r_B < r_A$ or $r_C < r_B$, and households with implausibly large interval differences between $r_B - r_A$ and $r_C - r_B$. "Implausible income" refers to households that report income outside $[0.5\times rmin, 2\times rmax]$, where *rmin* and *rmax* are replaced by r_A and r_C, respectively, if missing.

TABLE A.4. Colombia: bunching

	Round 1	Round 2
Threshold A (Lower)		
0%	1041	1036
50%	712	500
100%	83	23
Threshold B (Midpoint)		
0%	201	184
50%	806	762
100%	274	88
Threshold C (Higher)		
0%	85	36
50%	549	524
100%	1139	545
Available observations	7262	6295

Note. The table shows the number of respondents who reported 0%, 50% and 100% probabilities in each survey round.

TABLE A.5. Colombia: covariate balance (wave 1)

	(0)		(1)		(0)-(1)
Variable	*N*	Mean	*N*	Mean	
Number of adults	3670	2.73	2230	2.69	0.03
Number of female adults	3670	1.39	2230	1.38	0.01
Number of kids	3670	3.11	2230	3.25	-0.15^{***}
Log income	3670	12.74	2230	12.73	0.02
Rural household	3661	0.46	2225	0.45	0.01
Household head:					
Age	3657	44.30	2228	43.60	0.70^{**}
Some primary education	3612	0.43	2213	0.45	−0.02
Some secondary education	3612	0.15	2213	0.15	−0.01
Primary source of income:					
Laborer/employee	3464	0.29	2107	0.27	0.02
Domestic employee	3464	0.06	2107	0.05	0.00
Day laborer	3464	0.21	2107	0.23	−0.02
Self-employment	3464	0.39	2107	0.38	0.00
Partner in farm/plot	3464	0.06	2107	0.07	−0.01
Proportion of regular income	3664	0.79	2230	0.79	0.00
Health shocks	3670	0.13	2230	0.12	0.01
Other shocks	3670	0.14	2230	0.13	0.01

Note. (1) refers to observations in the final sample and (0) to available observations excluded from the final sample (just before "Balanced panel (robustness)" in Table A.3). First wave only. Robust standard errors; ***=.01, **=.05, *=.1.

TABLE A.6. Colombia: covariate balance (wave 2)

	(0)		(1)		(0)-(1)
Variable	*N*	Mean	*N*	Mean	
Number of adults	3118	2.80	2230	2.69	0.11***
Number of female adults	3118	1.43	2230	1.38	0.05**
Number of kids	3118	3.18	2230	3.25	−0.07
Log income	3118	12.73	2230	12.76	−0.02
Rural household	3106	0.47	2225	0.45	0.02
Household head:					
Age	3096	46.42	2216	45.39	1.03***
Some primary education	3027	0.45	2165	0.45	0.01
Some secondary education	3027	0.15	2165	0.18	−0.03***
Primary source of income:					
Laborer/employee	2990	0.31	2162	0.32	−0.01
Domestic employee	2990	0.05	2162	0.04	0.01**
Day laborer	2990	0.26	2162	0.26	0.00
Self-employment	2990	0.32	2162	0.32	0.00
Partner in farm/plot	2990	0.06	2162	0.06	0.00
Proportion of regular income	3115	0.90	2230	0.91	−0.02***
Health shocks	3118	0.15	2230	0.13	0.02*
Other shocks	3118	0.22	2230	0.18	0.04***

Note. (1) refers to observations in the final sample and (0) to available observations excluded from the final sample (just before "Balanced panel (robustness)" in Table A.3). Second wave only. Robust standard errors; ***=.01, **=.05, *=.1.

B Methodological appendix

Recall the flexible model in equation (9), which we reproduce below for convenience:

$$\ell_{jit} = \beta_0(r_{jit}) + \beta_1(r_{jit})\psi(y_{it}) + \beta_2(r_{jit})\eta_i + \varepsilon_{jit}. \tag{B.1}$$

We now provide further details on parameterization (Section B.1), estimation (Sections B.2 and B.3) and implementation (Section B.4).

B.1 Specification details

We view model (B.1) as a sequence of approximating parameter spaces — or sieves — for the nonparametric model in (3)-(4); see Chen (2007) for a technical review of the method of sieves.

In particular, we parameterize model (B.1) as

$$\ell_{jit} = \sum_{\tau=1}^{K_0} \beta_{0,\tau} h_\tau(r_{jit}) + \sum_{\tau=1}^{K_1}\sum_{\kappa=1}^{K_y} \beta_{1,\tau,\kappa} h_\tau(r_{jit}) g_\kappa(y_{it}) + \sum_{\tau=1}^{K_2} \beta_{2,\tau} h_\tau(r_{jit})\eta_i + \varepsilon_{jit}, \tag{B.2}$$

where $g_\kappa(y)$ are Hermite polymomials (we omit the constant term, i.e., g_1 is linear in y) and $h_\tau(r)$ are basis functions of natural cubic splines, with K_s knots ($K_s \geq 2$) for $s = \{0, 1, 2\}$.[24] For the ease of notation, we use L instead of K_s in the body of the paper, and explicitly refer to $\beta_0(\cdot)$, $\beta_1(\cdot)$ and/or $\beta_2(\cdot)$. We normalize $\beta_{0,1} = 0$ and

[24] Cubic natural splines are piece-wise cubic polynomials that are twice continuously differentiable and restricted to be linear beyond the boundary knots. Differentiability is crucial to compute densities and quantile-based measures of nonlinear persistence, as in (14). In particular, let τ_k for $k \in \{1, \ldots, K_s\}$ index the knots in increasing order, which we place at the $k/(K_s + 1)$th quantiles of the empirical distribution of r_{jit}. The following K_s basis functions can be used to represent the spline model:

$$\left[1, r_{jit}, d_1(r_{jit}) - d_{K_s-1}(r_{jit}), \ldots, d_{K_s-2}(r_{jit}) - d_{K_s-1}(r_{jit})\right],$$

so that $K_s = 2$ corresponds to a linear spline, and for $K_s > 2$ we have

$$d_k(r) = \frac{(r-\tau_k)^3 \mathbb{1}\{r \geq \tau_k\} - \left(r - \tau_{K_s}\right)^3 \mathbb{1}\left\{r \geq \tau_{K_s}\right\}}{\tau_{K_s} - \tau_k}$$

for $k \in \{1, \ldots, K_s - 2\}$.

$\beta_{2,1} = 1$ to accommodate the level and scale of the fixed effects η_i. This implies there are $K_0 + K_1 K_y + K_2 - 2$ target parameters in model (B.3). The baseline specification we use for nonlinear models in Sections 5.3 and 5.4 sets $K_0 = K_1 = 3$, $K_y = 1$ and $K_2 = 1$ (additive fixed effects) or $K_2 = 2$ (interacted fixed effects).

It is often useful to rewrite (B.3) in vector notation. If $h_{1:K_s}(r)$ is used for to indicate the $1 \times K_s$ array obtained by horizontal concatenation of the elements in $\{h_\tau(r)\}_{\tau=1}^{K_S}$, let us define

$$\begin{aligned} D_{jit} &= (h_{2:K_0}(r_{jit}), h_{1:K_1}(r_{jit}) g_1(y_{it}), \ldots, h_{1:K_1}(r_{jit}) g_{K_y}(y_{it}))' \\ \beta_{0,1} &= (\beta_{0,2:K_0}, \beta_{1,1:K_1,1}, \ldots, \beta_{1,1:K_1,K_y})', \end{aligned}$$

and similarly $H_{jit} = h_{2:K_2}(r_{jit})'$ and $\beta_2 = \beta'_{2,2:K_2}$. We then stack observations (t, j) vertically for each unit to obtain

$$\ell_i = D_i \beta_{0,1} + \left(1_{TJ} + H_i \beta_2\right) \eta_i + \varepsilon_i, \tag{B.3}$$

where 1_A is a vector of ones of size A and where $\ell_i = (\ell_{1i1}, \ell_{2i1}, \ell_{3i1}, \ell_{1i2}, \ell_{2i2}, \ell_{3i2})'$ and so on.

B.2 Estimation: additive fixed effects

Model (B.3) is then a series regression model on the sequence of parameter sets defined by $(K_{0,1,2}, K_y)$, with the additional twist that η_i is unobserved.

When η_i enters additively (set $K_2 = 1$), given the conditional mean assumption $E\left[\varepsilon_i \middle| r_i, y_i, \eta_i\right] = 0$, the model in (B.3) is a static fixed-effects regression that can be estimated using the within-group estimator. In other words, let $\tilde{\ell}_{jit} = \ell_{jit} - (TJ)^{-1} \sum_{(j,t)} \ell_{jit}$ denote variables in deviations with respect to means (recall that $T = 2$ and $J = 3$ here) and use $\tilde{\ell}_i$ for the corresponding vectors. Equation (B.3) then becomes $\tilde{\ell}_i = \tilde{D}_i \beta_{0,1} + \tilde{\varepsilon}_i$ and an estimate of $\beta_{0,1}$ can be obtained via least squares.

Penalization. When considering models where (K_0, K_1, K_y) grow large, regularized estimators might be attractive. We explore implementations that penalize the nonlinear sieve terms and might prove useful in richer setups where even more flexible specifications might be feasible. A simple implementation via a ridge

penalty $\lambda > 0$ allows us to maintain the simplicity of the estimation method and an explicit solution that recovers the linear model as $\lambda \to \infty$.

B.3 Estimation: interacted fixed effects

As noted in Section 3.5, treating $\{\eta_i\}_{i=1}^n$ as parameters to be estimated jointly with $\beta_{0,1}$ and β_2 results in an incidental parameters problem that precludes fixed-T consistent estimation. Here we generalize the linear instrumental variables (IV) strategy introduced in the main text, and discuss a method-of-moments approach in greater generality at the end.

Recall that we use $\tilde{\ell}_{jit} = \ell_{jit} - \bar{\ell}_i$ and so on as notation for variables in deviations with respect to means:

$$\begin{aligned}\tilde{\ell}_{jit} &= \tilde{D}_{jit}\beta_{0,1} + \tilde{H}_{jit}\eta_i\beta_2 + \tilde{\varepsilon}_{jit},\\ \bar{\ell}_i &= \bar{D}_i\beta_{0,1} + (1 + \bar{H}_i\beta_2)\eta_i + \bar{\varepsilon}_i.\end{aligned}$$

We consider the case $K_2 = 2$, so that effectively $H_{jit} = r_{jit}$. We look for linear transformations of the model that do not depend on η_i but still allow us to estimate β_2. Note that

$$\begin{aligned}H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit} &= H_{jit}\left(\bar{D}_i\beta_{0,1} + (1 + \bar{H}_i\beta_2)\eta_i + \bar{\varepsilon}_i\right) - \bar{H}_i\left(D_{jit}\beta_{0,1} + (1 + H_{jit}\beta_2)\eta_i + \varepsilon_{jit}\right)\\ &= \left(H_{jit}\bar{D}_i - \bar{H}_iD_{jit}\right)\beta_{0,1} + \tilde{H}_{jit}\eta_i + H_{jit}\bar{\varepsilon}_i - \bar{H}_i\varepsilon_{jit},\end{aligned}$$

where note that the first element in $H_{jit}\bar{D}_i - \bar{H}_iD_{jit}$ is zero. We solve for the term involving η_i and plug it in back in the model in deviations,

$$\tilde{\ell}_{jit} = \tilde{D}_{jit}\beta_{0,1} + \left(H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit}\right)\beta_2 - \left(H_{jit}\bar{D}_i + \bar{H}_iD_{jit}\right)\gamma + \xi_{jit} \tag{B.4}$$

where $\xi_{jit} = \tilde{\varepsilon}_{jit} - H_{jit}\beta_2\bar{\varepsilon}_i + \bar{H}_i\beta_2\varepsilon_{jit}$ and $\gamma = -\beta_{0,1}\beta_2$, a generalization of the simple model considered in equation (19) in Section (3.5.2). We need at least one instrument for $H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit}$. Note that

$$E\left[H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit}\middle|r_i, y_i\right] = \left(H_{jit}\bar{D}_i - \bar{H}_iD_{jit}\right)\beta_{0,1} + \tilde{H}_{jit}E\left[\eta_i\middle|r_i, y_i\right],$$

and thus any predictor of η_i (conditional on the included regressors) is possibly a valid instrument. We thus consider a set of $K_0 + K_1K_y - 1$ instruments given by

$Z_{jit} = \tilde{H}_{jit}\bar{D}_i$; this corresponds to five instruments in the baseline specification used in Section 5.4. We then propose to estimate (B.4) by TSLS.

Note that the restriction $\gamma = -\beta_{0,1}\beta_2$ does not need to be imposed for consistent estimation. If one is willing to impose the restrictions, this can be done expost via minimum distance estimation or exante via nonlinear GMM. We briefly explored the latter, which is exposed to similar numerical/convergence problems as those of the more general nonlinear method-of-moments estimator described below.

A method-of-moments estimator. The IV estimator developed here retains the simplicity and linearity of the within-group estimator even as we move on to more flexible models, and we have found it to be a reliable approach. A general method-of-moments approach for the parameters in equation (B.3) is as follows.

Let $B_i(\beta_2)$ and $Q_i(\beta_2)$ denote the generalized between- and within-group transformations, respectively, defined as

$$B_i(\beta_2) = \left(\left(1_{TJ} + H_i\beta_2\right)'\left(1_{TJ} + H_i\beta_2\right)\right)^{-1}\left(1_{TJ} + H_i\beta_2\right)',$$
$$Q_i(\beta_2) = I_{TJ} - \left(1_{TJ} + H_i\beta_2\right)B_i(\beta_2),$$

where I_A is the identity matrix of size $A \times A$. Back to equation (B.3), we note that the generalized within-group residuals are mean independent of the regressors:

$$E\left[Q_i(\beta_2)\left(\ell_i - D_i\beta_{0,1}\right)\middle| r_i, y_i\right] = 0.$$

A nonlinear GMM estimator inspired in Chamberlain (1992) and Arellano and Bonhomme (2012) is then available exploiting these conditional moment restrictions. In some sense, the IV strategy proposed above is a transparent way to finding such informative restrictions for the interacted fixed-effects term.

B.4 Additional details on implementation

Here we discuss how to we compute the objects of interest after estimation of the model parameters in equation (9) (reproduced here as equation (B.1)) and shed light on some additional details beyond those discussed in Section 3.5.

Standardizing the data. When estimating flexible models, we first standardize the data as follows:

$$\check{r}_{jit} = \frac{r_{jit} - \bar{r}}{\bar{\sigma}_r}$$
$$\check{y}_{it} = \frac{y_{it} - \bar{y}}{\bar{\sigma}_y},$$

where $\bar{x}$ and $\bar{\sigma}_x$ are measures of location and scale for variable x; we use the median and the IQR respectively in our implementation. This helps standardize the range of variation of the data across units. Importantly, we need to undo these transformations before reporting the final output: letting $\check{q}_{it}(\tau)$ denote the τth conditional quantile on the standardized data, we need to calculate

$$q_{it}(\tau) = \bar{r} + \bar{\sigma}_r \check{q}_{it}(\tau).$$

Estimation in growth rates. In Section 3.5, we note that redefining $s_{jit} = r_{jit} - y_{it}$ is equivalent to estimating predictive distributions for growth rates and argue that this is a convenient transformation. Note that we are still interested in the range of values of r (or $\check{r}$): this entails careful adjustment of the support grid of conditional distribution functions in implementation. On a related note, the fact that the argument of the conditional distribution now depends on y has to be taken into account when computing numerical derivatives below.

Details on computing quantile-based measures of dispersion, skewness and persistence. Given estimates $(\hat{\beta}'_{0,1}, \hat{\beta}'_2)'$, the target summaries in Section 3.4 can be computed in three steps:

1. Obtain predicted probabilities.

   Given reference conditioning values $(\bar{y}, \bar{\eta})$ (usually a quantile of interest) and for r in a given grid r_{grid}, we calculate fitted probabilities $\hat{p} = \hat{F}(r, \bar{y}, \bar{\eta})$, which we collect in $\hat{p}_{\text{rgrid}}$. When non-monotonic, we follow Chernozhukov et al. (2010) in sorting the original estimated curve into a monotone rearranged one.

2. Recover conditional quantiles.

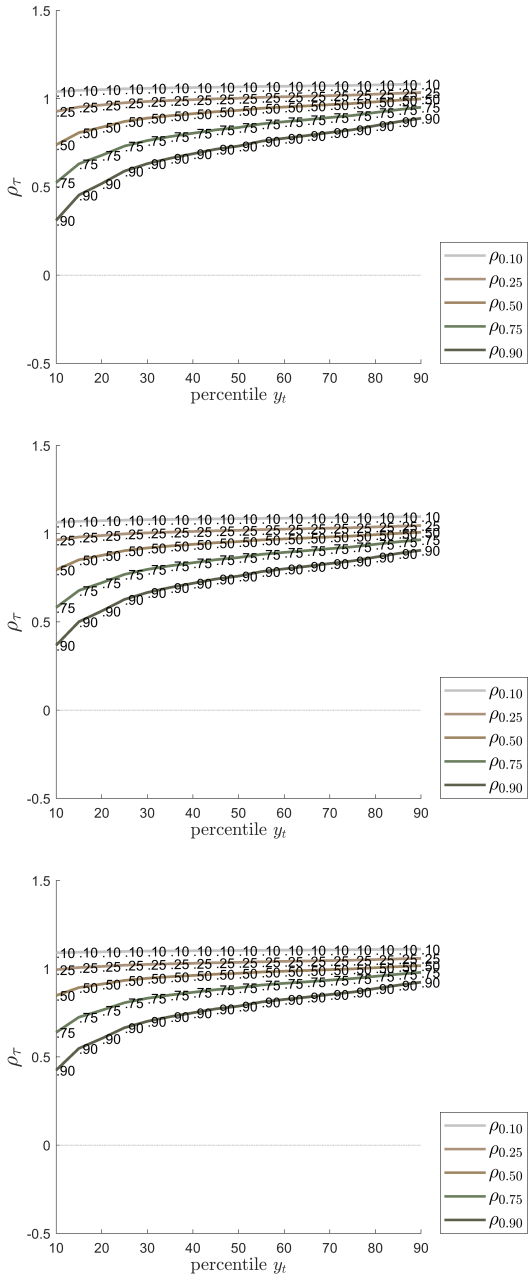
This involves inverting the fitted conditional distribution function to obtain conditional quantiles for a given τ, which we denote $\hat{r}(\tau)$; see also equation (10) in the main text. We resort to interpolation within $\hat{p}_{\text{rgrid}}$.[25]

3. Compute target summaries.

   Quantile-based measures of dispersion and skewness as in equations (12) and (13) are then readily available (note the comment above on standardizing the data). Regarding nonlinear persistence, we calculate the derivatives in equation (15) mumerically. Note that this requires recalculating predicted probabilities along the lines of step 1, so as to condition on $\hat{r}(\tau)$.

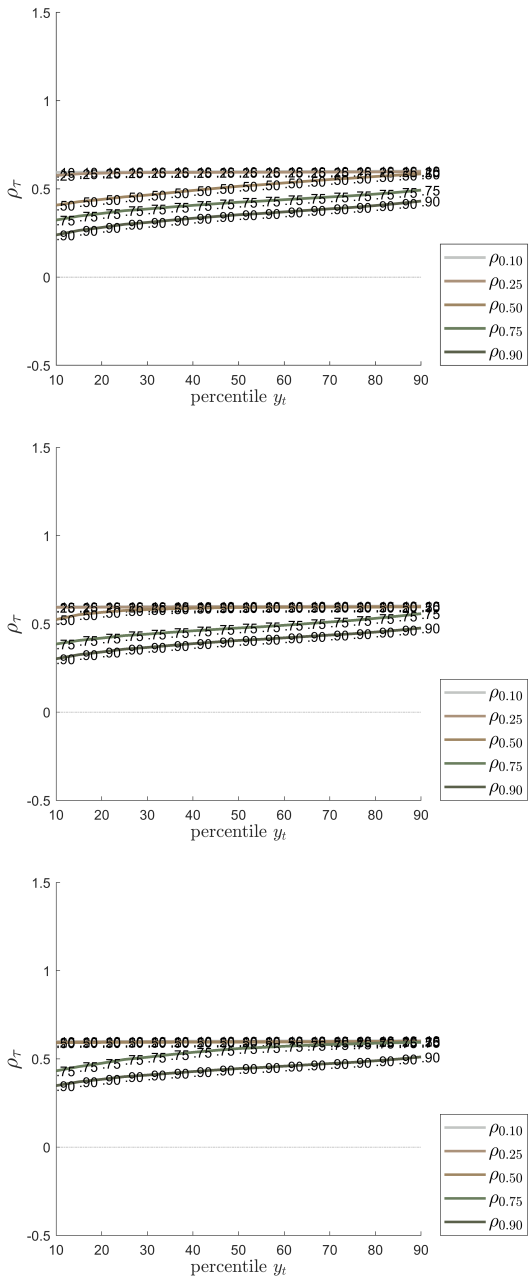
[25] Alternative methods are bracketing or root-finding alogorithms, which solve for $\hat{r}(\tau)$ in $\tau = \hat{F}(\hat{r}(\tau), \bar{y}, \bar{\eta})$. These are model-based approaches that impose the (estimated) logit structure, which is problematic when it is non-monotonic. In fact, these algorithms impose implicit rearrangement methods that are starting-value dependent.

C Documenting heterogeneity



Note. Reference values for $\bar{\eta}$ correspond to its 10th, 50th and 90th percentiles, respectively. See Figure C.3 for pointwise confidence bands.

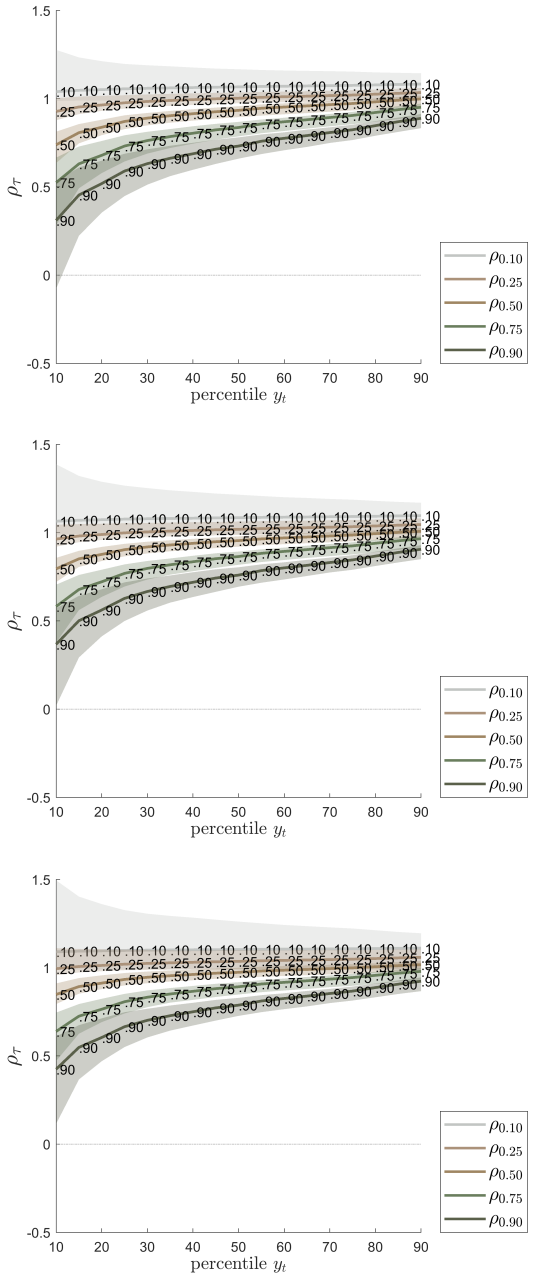
FIGURE C.1. India — nonlinear persistence at different reference values of η_i



Note. Reference values for $\bar{\eta}$ correspond to its 10th, 50th and 90th percentiles, respectively. See Figure C.4 for pointwise confidence bands.

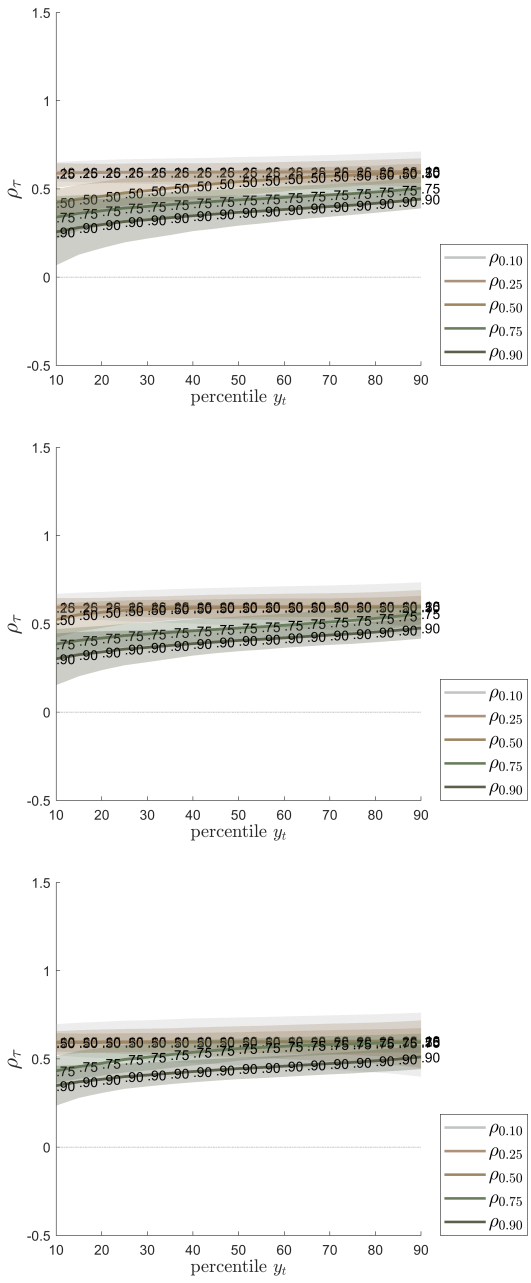
FIGURE C.2. Colombia — nonlinear persistence at different reference values of η_i

C.1 Confidence bands



Note. Reference values for $\bar{\eta}$ correspond to its 10th, 50th and 90th percentiles, respectively. 90% pointwise confidence bands; block bootstrap with 1000 repetitions.

FIGURE C.3. India — nonlinear persistence (with confidence bands)



Note. Reference values for $\bar{\eta}$ correspond to its 10th, 50th and 90th percentiles, respectively.

FIGURE C.4. Colombia — nonlinear persistence (with confidence bands)

D Robustness: sample selection and modeling choices

In Sections D.1 and D.2, we report counterparts to (a subset of) the empirical results reported in the main text when including elicited probabilities that equal zero or one (applying the transformation below) and those which violate strict monotonicity (two reported cumulative probabilities are equal for the same household). This entails minimal sample selection in the Indian data (see Table A.1) but is substantial in the Colombian data (see Table A.3). We find very similar results in both qualitative and quantitative terms for our target summary objects in both datasets. This is remarkable for the Colombian data, where using these transformations essentially imply doubling the sample size (from 2,230 to 4,420 unique households) and introducing substantial additional elicitation error (as measured by residual variances).

Keeping zero/one probabilities The logit transformation in equation (2) in Section 3 restricts observed, elicited probabilities p_{jit} to lie strictly between zero and one. We suggest here an alternative transformation — which still maps these probabilities to the real line — that allows us to keep these observations:

$$\ell_{jit} = \text{logit}\left(\check{p}_{jit}\right), \quad \check{p}_{jit} = \frac{p_{jit} + \frac{1}{2m}}{1 + \frac{J}{2m}}. \tag{D.1}$$

This is a generalization of the modified logit transformation of Cox and Snell (1970, p. 32) for binary data. In a context where p_{jit} are noisy measurements due to possible rounding and randomness in the elicitation process, the adjustment m can be interpreted as a measure of the accuracy of the elicitation such that $m = O\left(1/\sigma_\varepsilon^2\right)$. In particular, elicitations errors ε_{jit} in (3) can be seen as capturing sampling uncertainty from a hypothetical random sample of size m; see Arellano, Bonhomme, De Vera, Hospido, and Wei (2022, Online Appendix F) for additional details in the context of subjective expectations data and a Bayesian interpretation of (D.1).[26]

[26] Below, we set the regularization parameter m in equation (D.1) to $m = 10$ in both cases. Other reasonable choices lead to the same conclusions.

D.1 India: larger samples

	No FE	FE
ρ	0.96	0.93
	(0.94, 0.99)	(0.90, 0.96)
σ	0.58	0.33
	(0.53, 0.62)	(0.31, 0.35)
$IQR_{0.75}$	1.26	0.72
	(1.17, 1.36)	(0.67, 0.77)
$IQR_{0.90}$	2.53	1.44
	(2.34, 2.72)	(1.35, 1.53)
σ^2_η		0.18
		(0.15, 0.23)
σ^2_η village		0.12
		(0.11, 0.16)
σ^2_ε	1.05	0.98
	(1.03, 1.08)	(0.95, 1.02)

Note. The table reports results for the linear model in (7) using the data for India, without fixed effects (and a common intercept) and with fixed effects. Results correspond to the sample that includes reported zero/one probabilities (see Section D) and those which violate strict monotonicitiy; this is the counterpart to Table 3 in the main text. Specifications include year (survey round) dummies in both cases. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE D.1. India — linear model (robustness sample)

	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	0.83	0.63	0.54
	(0.75, 0.93)	(0.57, 0.66)	(0.48, 0.58)
$IQR_{0.90}$	1.69	1.29	1.10
	(1.56, 1.91)	(1.21, 1.37)	(0.99, 1.21)
$SK_{0.90}$	−0.04	−0.11	−0.15
	(−0.16, 0.04)	(−0.21, −0.05)	(−0.29, −0.05)
$\rho_{\tau 0.25}$	0.97	1.02	1.04
	(0.92, 1.04)	(0.98, 1.06)	(0.99, 1.08)
$\rho_{\tau 0.50}$	0.81	0.95	1.00
	(0.74, 0.87)	(0.92, 0.98)	(0.97, 1.02)
$\rho_{\tau 0.75}$	0.59	0.86	0.95
	(0.42, 0.72)	(0.82, 0.89)	(0.91, 0.98)
σ^2_η		0.19	
		(0.16, 0.23)	
σ^2_η village		0.11	
		(0.11, 0.16)	
σ^2_ε		0.95	
		(0.91, 0.99)	

Note. The table reports results for India for the flexible model with additive fixed effects in (20). Results correspond to the sample that includes reported zero/one probabilities (see Section D) and those which violate strict monotonicitiy; this is the counterpart to Table 9 in the main text. We also include year (survey round) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE D.2. India — flexible model (additive fixed effects)

D.2 Colombia: larger samples

	No FE	FE
ρ	0.72	0.51
	(0.69, 0.75)	(0.48, 0.54)
σ	0.82	0.56
	(0.80, 0.85)	(0.55, 0.58)
$IQR_{0.75}$	1.80	1.24
	(1.75, 1.86)	(1.21, 1.27)
$IQR_{0.90}$	3.61	2.48
	(3.49, 3.72)	(2.41, 2.55)
σ^2_η		0.54
		(0.51, 0.58)
σ^2_η village		0.10
		(0.10, 0.13)
σ^2_ε	1.75	1.34
	(1.71, 1.78)	(1.30, 1.37)

Note. The table reports results for the linear model in (7) using the data for Colombia, without fixed effects (and a common intercept) and with fixed effects. Results correspond to the sample that includes reported zero/one probabilities (see Section D) and those which violate strict monotonicitiy; this is the counterpart to Table 4 in the main text. Specifications include year (survey round) and month (interview) dummies in both cases. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE D.3. Colombia — linear model (robustness sample)

	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	1.36	1.18	1.10
	(1.31, 1.42)	(1.14, 1.21)	(1.06, 1.14)
$IQR_{0.90}$	2.67	2.37	2.22
	(2.57, 2.78)	(2.31, 2.44)	(2.15, 2.29)
$SK_{0.90}$	0.08	0.05	0.01
	(0.04, 0.13)	(0.01, 0.08)	(−0.02, 0.04)
$\rho_{\tau 0.25}$	0.60	0.63	0.65
	(0.55, 0.64)	(0.59, 0.67)	(0.58, 0.70)
$\rho_{\tau 0.50}$	0.45	0.60	0.63
	(0.41, 0.51)	(0.56, 0.63)	(0.59, 0.66)
$\rho_{\tau 0.75}$	0.34	0.50	0.59
	(0.28, 0.40)	(0.46, 0.53)	(0.55, 0.62)
σ^2_η		0.53	
		(0.50, 0.57)	
σ^2_η village		0.10	
		(0.10, 0.13)	
σ^2_ε		1.33	
		(1.29, 1.36)	

Note. The table reports results for Colombia for the flexible model with additive fixed effects in (20). Results correspond to the sample that includes reported zero/one probabilities (see Section D) and those which violate strict monotonicitiy; this is the counterpart to Table 10 in the main text. We also include year (survey round) and month (interview) dummies. In parenthesis we report 90% block bootstrap CI (1000 repetitions).

TABLE D.4. Colombia — flexible model (additive fixed effects)

E Questionnaires

Figure E.1 reports the original questionnaire for India. The questions on elicitation of subjective expectations follow those on income and income components and correspond to section 6 of the household survey.[27] Figure E.2 shows the original questionnaire for Colombia (in Spanish).

INTERVIEWER: Add all income sources in the shaded column to calculate yearly income of the household.

5.	READ OUT CALCULATED YEARLY INCOME and ask: Is this a typical yearly income for your household?	1. yes 2. no, it is higher than typical 3. no, it is lower	
6.	IF NO: What would be a typical yearly income for your household?	(Rs.)	

IF ONLY INCOME SOURCE IS FROM DAIRY ACTIVITY (7) >> GO TO SECTION 7. ELSE, go on to question 7.

7.	Imagine that you have a very good year, every member of working age in the household managed to have work, and there were no droughts or anything the like. What would be the maximum amount of income your household would receive in such a situation in one year?	Y	(Rs.)	
8.	Now imagine the total opposite: the harvest is bad, animals get sick, finding work is not possible. What would be the yearly income of your household in such a situation?	X	(Rs.)	

INTERVIEWER: Calculate the following values:

Expected Income (threshold B):	B = (X+Y)/2	
Threshold A:	A = (B+X)/2	
Threshold C:	C = (B+Y)/2	

INTERVIEWER: Explain the rainfall question to the respondent (See extra Sheet)

R.1	So, what do you think how likely it is that it will rain *tomorrow*?	
R.2	So, what do you think how likely it is that it will rain within the *coming week*?	
R.3	So, what do you think how likely it is that it will rain within the *coming month*?	

9.	How likely do you think it is that your yearly income in the coming year will be higher than _____(A) Rupees?	
10.	How likely do you think it is that your yearly income in the coming year will be higher than _____(B) Rupees?	
12.	How likely do you think it is that your yearly income in the coming year will be higher than _____(C) Rupees?	

FIGURE E.1. India — questionnaire

[27] To be precise, this is the second-round version of the questionnaire. In the first-round version, five households were instead asked about monthly — rather than yearly — income. Importantly, the wording of the questions is unchanged, and the data included an identifier for these households, which are not part of the original sample in Table A.1.

FIGURE E.2. Colombia — questionnaire