Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Estimating flexible income processes from subjective expectations data: evidence from India and Colombia

Manuel Arellano, Orazio Attanasio, Sam Crossman, Victor Sancibrian

Working Paper n. 733

This Version: July 2026

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
http://www.igier.unibocconi.it

Estimating flexible income processes from subjective expectations data: evidence from India and Colombia*

Manuel Arellano† Orazio Attanasio‡ Sam Crossman§ Víctor Sancibrián¶

July 2026

Abstract

We develop a methodology for modeling perceived 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 subjective persistence. We then use two longitudinal surveys collected in rural India and rural Colombia to explore the nature of perceived 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.

Keywords: Subjective expectations, income processes, nonlinear persistence.
JEL classification: C23, C81, D15.

*We thank Alberto Abadie, Aureo De Paula, Laura Liu, Wilbert van der Klaauw, Johannes Wohlfart, a co-editor, three anonymous referees, 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 Indian data used in this paper were collected in 2008 and 2009 for the evaluation of a microfinance program. Britta Augsburg organized the data collection and provided useful information on the survey and the environment where the data were collected. Yahu Cong, José M. Cueto and Stefano Graziosi 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. A big part of this work was done while Sancibrián was at CEMFI. The paper contents do not reflect the views of the UK Government Economic Service.

†CEMFI: arellano@cemfi.es
‡Yale University and NBER: orazio.attanasio@yale.edu
§UK Government Economic Service: samuelcrossman@gmail.com
¶Bocconi University and IGIER: victor.sancibrian@unibocconi.it

1 Introduction

Risk is important for household choices. 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. Analogous considerations apply to labor supply and employment decisions. *Perceived* uncertainty is key for many decisions. Its heterogeneity across households might therefore be a key factor in understanding heterogeneous choices and inequality at different levels.

Conventional approaches to identify the stochastic process of uncertain variables, such as those discussed in Meghir and Pistaferri (2011), Arellano (2014), and Altonji, Hynsjö, and Vidangos (2023), are indirect, relying on statistical models of the dynamics of the realized variables and/or models of choice. Under rational expectations, the probability distribution of income and its dynamics, as perceived by households, may coincide with the probability distribution of the dynamics of realized income, but this need not be the case. An alternative *direct* approach is to rely on subjective expectations survey data. In this paper, we develop a methodology for modeling household income processes using probabilistic 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 apply our methodology to subjective expectations survey data from India and Colombia. Learning about the nature of income uncertainty is particularly important in developing economies, where income uncertainty and volatility tend to be higher.

Approaches based on realized income and those based on subjective expectations share a common objective: understanding the uncertainty that is relevant for agents' economic decisions. The literature using realized outcomes has typically assumed, implicitly or explicitly, that this uncertainty can be recovered from realizations alone. While this often relies on some form of rational expectations, it need not be enough. For example, agents may be rational but possess information that cannot be inferred from realized outcomes alone. We therefore do not regard the dynamics recovered from realized outcomes as inherently more "true" or "objective" than those implied by subjective expectations. In this paper, we emphasize the value of directly eliciting perceived income dynamics, but we do not take a stand on whether they coincide with those implied by realized outcomes, or on the sources of any differences. Those are important questions, but they lie beyond the scope of this paper. Our contribution is methodological and empirical: we propose a new way of identifying economically relevant uncertainty, complementing rather than replacing existing approaches.

Subjective expectations data have been around for some time, although they were initially received with some skepticism. However, the available 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, 2011b; Kosar and O'Dea, 2023; Fuster and Zafar, 2023), and much progress has been made in understanding the implications of different elicitation methods in both developed and developing countries (Attanasio, 2009; Delavande, 2023).

In contrast to the vast literature studying income processes using data on realizations, there is relatively little work on perceived income dynamics. Some of the early work on subjective income expectations is due to Dominitz and Manski (1997b), who used survey expectations to fit respondent-specific parametric distributions and compared them to those implied by the income processes estimated using realizations in Hall and Mishkin (1982). Jappelli and Pistaferri (2000) and Pistaferri (2003) used Italian data on subjective expectations to test for excess sensitivity of consumption to income shocks and to decompose income processes into different components.

In recent years there has been a surge of interest in this area. Subjective income and earnings expectations have been used to study income processes (Attanasio and Augsburg, 2016), to separately identify transitory and persistent income shocks (Attanasio, Kovacs, and Molnar, 2020), to develop tests of rational expectations (D'Haultfoeuille, Gaillac, and Maurel, 2021), and to characterize the earnings and employment risk that workers face through the life cycle (Caplin, Gregory, Lee, Leth-Petersen, and Sæverud, 2024; Kosar and van der Klaauw, 2025), to cite a few examples. Bachmann, Topa, and van der Klaauw (2023) provide a thorough discussion of the collection, study and use of expectations data in economics.

In this paper, we develop a methodology for modeling income processes that leverages the availability of probabilistic expectations of future income across different households and points in the distribution. 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-odds 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-modeled 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 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, Fella, 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.

After using the log-linear model as a starting point that conveys the main ideas of our approach, we 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-odds 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 a Colombian and an Indian survey, combining expectations with actual income data. In fact, the combination of the two is essential to make progress: even under rational expectations, the objects of interest are not identified with two waves and data on income realizations alone. In both surveys, income expectations were collected using the 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.

We find low persistence for large positive shocks at the bottom of the income distribution, but not for large negative shocks at the top.[1] We argue that the nonlinear persistence we find for the poorest households is consistent with a poverty trap interpretation: poor households hit by bad shocks expect their low income to perpetuate for longer, but instead they anticipate that large, positive shocks can persistently weaken the weight of the past. 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.

[1] These 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).

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 Appendix contains technical material, and additional results are relegated to the Supplemental Appendix.[2]

2 Eliciting subjective expectations

Designing questionnaires to elicit respondents' perceived probability distributions is challenging, with ongoing debates about how to elicit conditional or unconditional probabilities. As a result, key design choices must be made throughout the process. This section briefly reviews the main issues in the literature and outlines the approach used to elicit subjective expectations in our two surveys, which shared similar methods and questionnaire designs. While not novel, describing them helps situate our choices within existing alternatives.

2.1 Anchoring subjective expectations

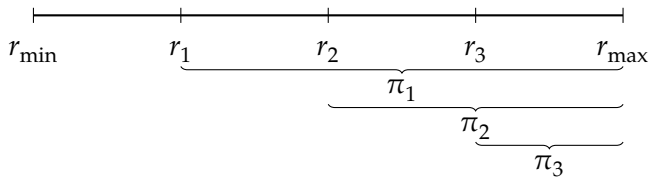
A central challenge in designing subjective expectation questions is defining an appropriate anchor and metric for the variable whose probability distribution is to be elicited. In the case of future income, two main approaches have been used. Some surveys use current income as an implicit anchor and ask respondents about the probability of various percentage changes relative to this baseline. Others ask respondents to specify a range of possible future values — typically the minimum and maximum — which are then used to construct intervals over which probabilities are elicited. This method, developed by Dominitz and Manski (1996, 1997a,b), has been widely adopted in surveys in both developed and developing countries, including the Indian and Colombian datasets used here. The questionnaires are reported in Supplemental Appendix E.

Morgan and Henrion (1990) note that asking respondents first for the minimum and maximum of possible outcomes helps reduce overconfidence — by encouraging consideration of the full range of outcomes — and anchoring, since respondents define their own reference points rather than relying on interviewer-provided figures. Using current income as an anchor can also be problematic when it is unusually low or high, since the resulting distribution may not be representative. The minimum/maximum approach avoids this issue by allowing respondents to recenter their expectations based on their own information. However, it remains unclear whether the elicited minimum and maximum values truly reflect the extremes of their subjective distribution or rather some arbitrary low and high percentiles.

[2]The Supplemental Appendix can be accessed at https://sancibrian-v.github.io/files/AACS-subjective-supp.pdf.

2.2 Eliciting subjective probability distributions

Having recorded the minimum and maximum possible future incomes, which we denote as $r_{\min}$ and $r_{\max}$, respectively, interviewers elicit several intermediate points on respondents' subjective cumulative distribution function (*cdf*) of future household income. Using a simple pre-specified algorithm, three (respondent-specific) equally spaced points $\{r_1, r_2, r_3\}$ are computed within the range $[r_{\min}, r_{\max}]$, and respondents report the probabilities $\{\pi_1, \pi_2, \pi_3\}$ that next-period income will exceed each threshold. This procedure, illustrated in Figure 1, was applied in both the Indian and Colombian surveys.



Notes: Respondents report minimum and maximum future income values, $r_{\min}$ and $r_{\max}$. Thresholds are defined as $r_2 = (r_{\min} + r_{\max})/2$, $r_1 = (r_{\min} + r_2)/2$, and $r_3 = (r_2 + r_{\max})/2$. Respondents then report probabilities π_1, π_2, and π_3.

FIGURE 1. Elicitation of the subjective income distribution

A common concern with this approach is that respondents — especially those with limited formal education — may lack familiarity with probabilistic reasoning. To mitigate this potential problem, both surveys included a short priming module introducing the concept of probability and conditional probability through examples of uncertain events, such as rainfall. During the priming, questions were structured to reinforce consistency. For instance, the probability of rain in the next seven days should be at least as high as for tomorrow; if respondents gave inconsistent answers to these questions, they were made to notice the inconsistency. No corrections were made during the main income expectation module. As a result, some inconsistencies remain, which we examine in Section 4.

To aid understanding, a visual "probability ruler" from 0 to 100 was used, allowing respondents to point to values indicating their degree of certainty in future uncertain events. While effective in these settings, alternative visual tools may be preferable elsewhere — for instance, allocating stones or tokens into bins, or eliciting probability densities directly (Delavande and Rohwedder, 2008; Delavande, Giné, and McKenzie, 2011a).

Finally, the choice of the number of elicited points J involves a trade-off. A larger J increases information about the subjective *cdf* but raises respondent burden and risks reducing data quality. In practice, most surveys—including ours—use $J = 3$. In what follows, we make no assumption that the reported minima and maxima represent true bounds of the subjective income distribution; they serve solely to define an individualized range over which the *cdf* is elicited.

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 realizations. 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 income process models we present 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 substantially weaker than those used when estimating such models with actual income realizations. 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 fact that, as in other surveys used in this growing literature, respondents are asked not only for point expectations of a future variable but also for some points on that variable's *cdf*. The availability of such data allows us to fit a model for the conditional *cdf* directly to the observed individual subjective probabilities. Let a household's subjective (conditional) cumulative probability 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 $p_{jit} = 1 - \pi_{jit}$ 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 $\text{logit}(p) = \ln[p/(1-p)]$. 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 classical measurement error, in the sense that it is mean independent of the variables in 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 might be 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.

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 of the nature and sources of income, x_{it}, such as the number of earners in the household, as well as household demographics. It is important to account for a comprehensive set of observables, since omitting relevant predictors (beyond fixed effects) would lead to potential endogeneity issues. 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 is intended to encompass both household-level characteristics and geographical (say, village-level) characteristics. We therefore assume that $I_{it} = \left(y_{it}, x_{it}, r_{1it}, \ldots, r_{Jit}, \alpha_i\right)$. The individual effect α_i may be arbitrarily correlated with the other elements in the information set.

[3] Notice that this transformation rules out observations with $p_{jit} = 0$ or $p_{jit} = 1$. In Supplemental 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.

Identification and data requirements. 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 that we 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 r_{jit}, which represents the cutoff points at which the subjective *cdf* is observed. The function g depends on the specific model of income dynamics one uses; below we discuss different models. Substituting equation (4) into equation (3) we obtain:

$$\ell_{jit} = g\left(r_{jit}, y_{it}, x_{it}, \alpha_i\right) + \varepsilon_{jit} \qquad (i = 1, \ldots, n; j = 1, 2, 3; t = 1, 2)$$

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 function g and the conditional distribution of α_i in model (3) and (4) can be established using the arguments in Evdokimov (2010), under the assumption that the additively separable disturbance terms are conditionally independent over time, and independent of the individual effect α_i.

Intuitively, this argument of nonparametric identification relies on sufficient variation in r_{jit} over both units and points of the distribution. In other words, identification of the model might still be possible if $J = 1$ but there is sufficient variation across units and time in the elicited point. An example where this fails is a survey question where the elicited support points $\{r_{jt}\}_{j=1}^{J}$ are the same for all units, in which case the conditional distribution is only identified at those points. From this point of view, eliciting a larger number of probabilities J is beneficial, but it also tends to make survey questions more demanding and might lead to larger elicitation errors. Beyond this, our methods are also immediately applicable to alternative data collection strategies, such as those that elicit probabilities on intervals of the form $\{r_0 < y_{i,t+1} \leq r_1\}$, and those where current income is used instead as an anchor.

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

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

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, $V_{i,t}$ captures 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 decompose the stochastic component of income into persistent and transitory shocks. In standard persistent–transitory models, consumers are assumed to observe these two components as separate state variables, even though they remain unobserved to the econometrician. When conditioning on current income, as we do, this framework gives rise to a measurement-error problem that can in principle be addressed using instrumental variables. However, such an approach is not feasible in a two-wave panel like ours. We therefore do not explicitly pursue this decomposition in our application.

Instead, we estimate income processes that are considerably more flexible than the simple log-linear specification introduced above. First, we allow for heterogeneity in income dynamics by interacting current income with indicators of the nature and sources of income, an approach whose motivation is closely related to the motivation underlying unobserved-component models. Moreover, we consider a class of processes with nonlinear persistence, such as those studied by Arellano et al. (2017). In these models, log income $y_{i,t}$ follows:

$$y_{it} = Q_t(y_{i,t-1}, u_{it}), \quad (u_{it} \mid y_{i,t-1}, y_{i,t-2}, \ldots) \sim \text{Uniform}(0,1),\ t = 2, \ldots, T.$$

Below, we discuss how to approximate the function $g(\cdot)$ in equation (4) and map it into flexible models of this type.

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 with zero mean and unit variance and to be independent of y_{it} and α_i. The corresponding conditional *cdf* is then

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

where $\tilde{\sigma} = \sigma\sqrt{3}/\pi$ and $\Lambda(x) = (1 + \exp(-x))^{-1}$ 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 $\beta_0 = 1/\tilde{\sigma}$, $\beta_1 = -\rho/\tilde{\sigma}$ and $\eta_i = -\alpha_i/\tilde{\sigma}$. 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.

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

[5] We include time dummies in all models that we consider (year fixed effects in India and year and month fixed effects in Colombia), but omit them from explicit formulas for notational simplicity. When reporting results for nonlinear models, we condition on the first year (or year and month).

[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.

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 approach based on realizations would need at least three. In other words, the combination of expectations and realizations is what allows us to explore the nature of income dynamics in our empirical illustrations.

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 relatively simple way to relax the restriction that persistence is common to all households is to allow for interactions of current income with relevant time-varying observable household characteristics x_{it}, which extends equation (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.

3.3 Flexible income processes

We now consider generalizations of the linear model to approximate conditional predictive distributions of the form discussed in equation (4). In particular, consider the following flexible 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]

[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 Appendix A effectively incorporates such interactions.

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 $\mathbb{1}\{y_{i,t+1} < r\}$ 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 $\beta_s(\cdot)$ and $\psi(\cdot)$ need to be parameterized. We specify $\beta_s(\cdot)$ as natural cubic splines with K knots, and $\psi(\cdot)$ as a vector of low-order Hermite polynomials in standardized current income. These flexible parametric models are taken as approximations to the unspecified function $g(\cdot)$, and the logistic assumption effectively remains just a convenient transformation.[9]

Despite this flexibility, estimation remains tractable. In specifications with $\beta_2(r_{jit}) = 1$, the model remains a static linear panel model, and the simple within-group estimator can be used to recover the parameters. In models with interacted fixed effects, we propose a simple instrumental-variables estimator that exploits the predictive content of (averages of) r_{jit} and y_{it} for fixed effects η_i. We present further details concerning implementation and estimation in Appendix A.

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 (6), the conditional quantile is defined as

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

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

A standard measure of dispersion is the interquantile range:

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

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

[9] In our implementation, we introduce additional flexibility by instead approximating predictive distributions for income growth, that is, $P\left(\Delta y_{i,t+1} < s_{jit} \middle| I_{it}\right)$ for $s_{jit} = r_{jit} - y_{it}$. Further details are discussed in Appendix A.

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

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

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:

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

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, in the linear case, 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 were 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.

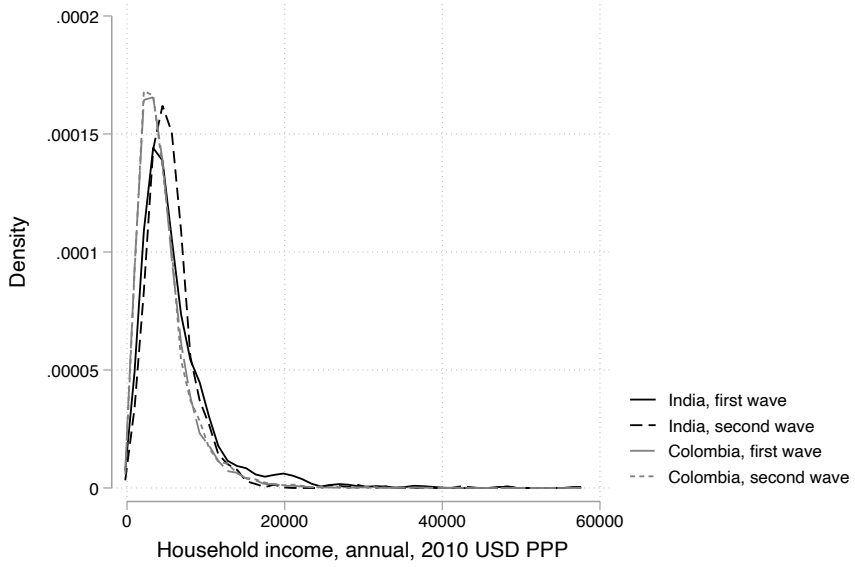
In our analysis, we do not rely on *realized future outcomes* and *realized future shocks*. Instead, by using data on subjective expectations, we leverage the subjective conditional probability distribution of future outcomes. This approach allows us to examine the impact of *potential (subjective) shocks* on potential future outcomes, highlighting the additional information that expectations data provide when combined with current income realizations.

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 poor rural populations.

In Figure 2, we plot the distribution of reported total household income across countries and survey waves. In the analysis, we work with yearly and monthly household income in 2010 constant prices in India and Colombia, respectively, but here we convert both measures to annual 2010 PPP USD for comparability. While the contexts we are studying are very distinct, the two distributions are remarkably similar, indicating that we are studying comparably poor populations. Average annual household income in the sample is $5,980 for India and $4,650 for Colombia, with a standard deviation of $4,664 and $3,489, respectively.



Notes: Distribution of household income in the two study populations, expressed 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 it refers to the following month.

4.1 India

The Indian data 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 cattle); see Augsburg (2009) and Attanasio and Augsburg (2016) for further details. A typical household in our sample

has five members and a male, 45-year-old household head. Most households belong to the "Other Backward Classes", a term used by the Government of India to denote disadvantaged castes. 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 did not receive 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 laborers (64%).

TABLE 1. Income shocks and sources, India

	≤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

Notes: $N = 770$ households. Relative frequencies of household income components for India, pooled across waves. The first row reports the share of households with up to two, three, and four or more income sources. The next three rows show the share of current income from farm-related activities by income-source group. The final four rows report the relative frequency of major negative shocks experienced during the previous year, by income-source group (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).

In Table 1, we present descriptive information on income sources (defined as the number of activities from which the household generates income, such as farming, agricultural labor, relief work, crafts, etc.) and major negative shocks experienced during the previous year, such as health shocks or crop loss due to natural disasters, 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.

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 up 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. Additional details are reported in Table A.1.[10]

About a quarter of these households were clients of a microfinance institution (MFI) and had received loans in 2008 for livestock investment. The remaining households were either residing in the same villages or in villages the MFI considered targeting in the future. The data were 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.

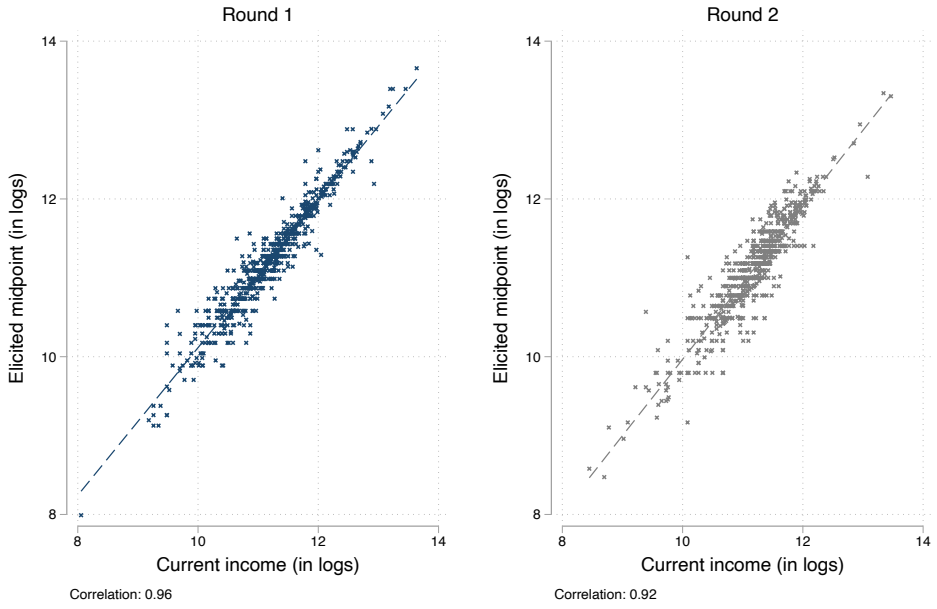
In both survey waves, the interviewers — who visited the respondents in their homes — elicited information about points on the respondents' subjective household income distribution. The technique discussed in Section 2 was used after explaining the approach in detail, practicing with rainfall questions, 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 that 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 otherwise made use of the entire range, although some bunching at multiples of 5 was observed (possibly because these were indicated on the ruler), and (v) expectations correlate sensibly with household characteristics. We report further details in Supplemental Appendix A.1.

Figure 3 plots the elicited midpoint of the predictive distribution (computed as the average of min and max future expected income) against realized current income, and shows an extremely high cross-sectional correlation between households' 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.

Figure 4 provides a summary of the distribution of elicited probabilities by threshold and survey round; the data for India are displayed in panel (a). 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. In 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.

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



Notes: The solid line shows the linear regression fit. Annual household income, 2010 INR.

FIGURE 3. Current income and reported midpoint, India

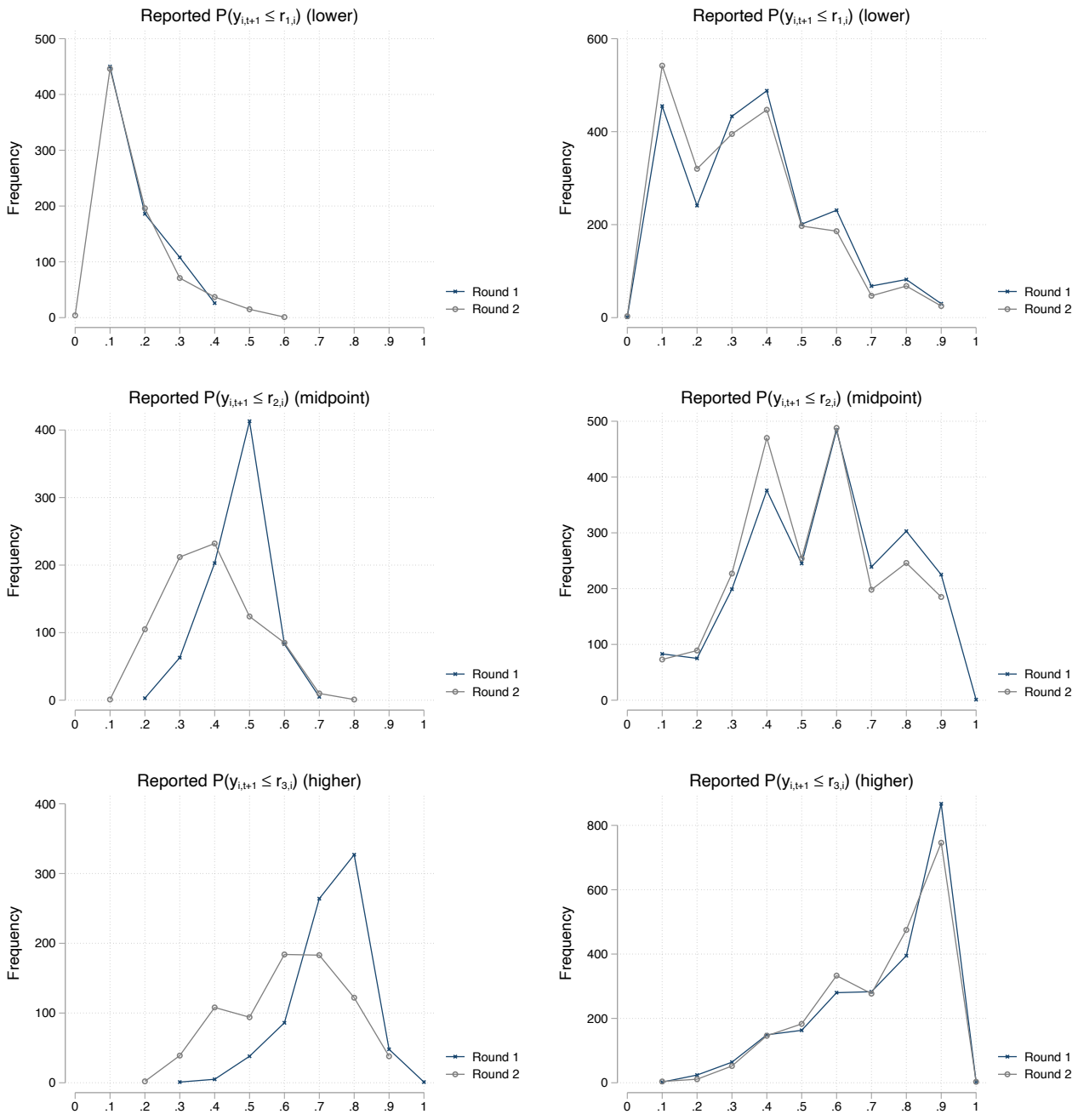
4.2 Colombia

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 improved nutrition, health, and education in rural Colombia. As with 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.[11]

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

Survey respondents are predominantly female (65%). Just over half (54%) are household heads,

[11] 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).



(a) India (b) Colombia

Notes: Frequencies of elicited probabilities (rounded to the nearest tenth) at each threshold, by survey round. See Figure 1 for a description of the elicitation process.

FIGURE 4. Distribution of elicited probabilities by threshold

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 not completed primary

TABLE 2. Income shocks and providers, Colombia

	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

Notes: $N = 2230$ households. Relative frequencies of household income components for Colombia, pooled across waves. The first row reports the share of households with one, two, or three or more earners in the previous month. The next three rows show the share of current income from regular (rather than occasional) sources by number of earners; "more than 75% regular" excludes 75% and 100%. The final three rows report the relative frequency of major negative shocks experienced during the previous year, by number-of-earners group. A small fraction of households (about 2.5%) report both health and other shocks, so absolute frequencies exceed the reported categories.

education, with an average of 3.5 years of schooling. Income is predominantly earned in the form of labor income, and most individuals tend to be informally employed (93%). About half of the working 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).[12]

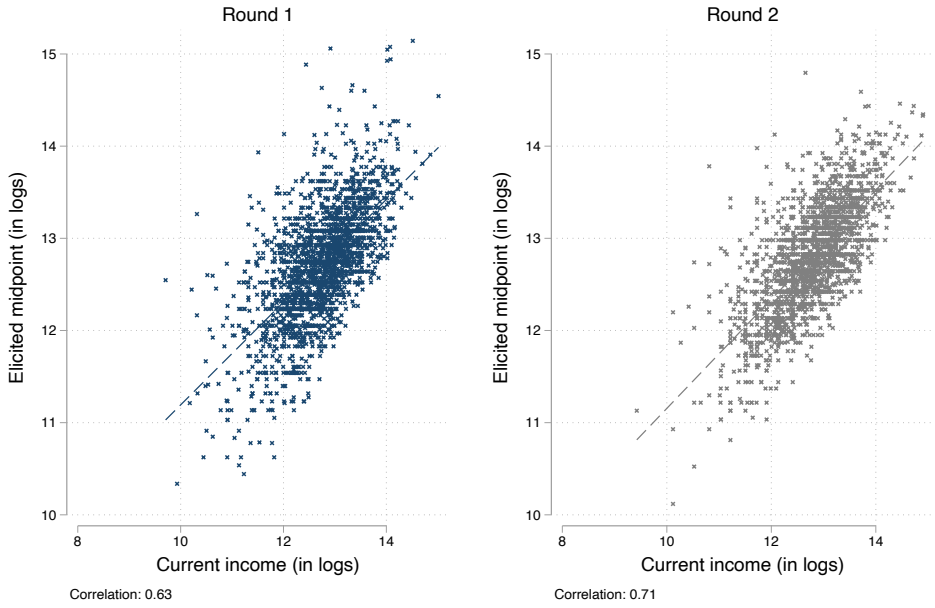
In Table 2, as for the Indian data, we provide descriptive statistics of time-varying characteristics related to income sources and shocks. We divide households into three categories according to the number of household members who report a source of income during the previous month: one, two or three 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 labor, 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 income categories, similar to the pattern observed in India.

Elicitation of expectations was conducted in a similar fashion to that used in India; see again Section 2 for a description of the elicitation approach. The main difference relative to the Indian expectations survey is the *horizon* of future income. In the India survey, future income refers to the following year, while in the Colombian survey, future income refers to the following month.

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

Figure 5 displays a high positive correlation between the midpoint of the reported min and max future income and current income in both survey rounds. This relationship is somewhat weaker than in the Indian context, in line with the higher risk and lower persistence documented below, and with the fact that expectations here refer to a much shorter time span.

The right panel of Figure 4 summarizes the distribution of elicited probabilities by threshold and survey round, as with the Indian survey. The distribution of elicited probabilities for each of the three cutoff points exhibits more heterogeneity than in the Indian sample, in particular for the first two cutoffs.



Notes: The solid line shows the linear regression fit. Monthly household income, 2010 COP.

FIGURE 5. Current income and reported midpoint, Colombia

Since the subjective expectations data have not been used before, we provide a detailed analysis and validation in Supplemental Appendix A.2. Overall, the elicitation of subjective expectations was less precise in the Colombian data. We find a substantially higher rate of logical response errors, although still within reasonable ranges (for instance, around 4% of households report distributions that violate monotonicity), and a substantial fraction of households have either missing income data or missing elicited minimum/maximum ranges and probabilities. These validation and sample selection steps leave us with a (balanced) panel sample of $N = 2,230$ households; see Table A.1 in Supplemental Appendix A.2 for a detailed step-by-step analysis. Nonetheless, Tables A.3 and A.4 show that these decisions do not imply strong selection based on observable characteristics,[13] and

[13]Households in the final sample tend to have fewer adults, and household heads are slightly younger (by around a

we find very similar results when using an extended sample of 4,420 households, which includes households that report probabilities equal to zero or one or those which are not consistent with a strictly increasing *cdf* (see Supplemental Appendix D.2 for details).

5 Results

In this section, we report the results we obtain for both countries when estimating various specifications of the income process. In Section 5.1, we start with a linear AR(1) process with time effects and consider versions with and without fixed effects. In Section 5.2, we consider nonlinear processes of the type introduced in Section 3.3. In such models, all the features of the distribution vary across units and over time. Finally, in Section 5.3, we report results for models with interacted fixed effects.

5.1 Linear models

In this sub-section, 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

The estimates of the linear model using the Indian data are reported in Table 3(a). 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.[14]

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 substantial at 1.01, reflected in the large values of the reported interquantile ranges. This estimate can be interpreted as a measure of risk, identified from the self-reported range of variation of future income and elicited probabilities. The measurement error 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 deviation of the income-process innovations is reduced from 1.01 to 0.57.

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.

[14] Consistent estimators of variance components are obtained following Arellano and Bonhomme (2012).

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 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 a 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 15 percentage points.[15] 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. In contrast, the flexible models we consider in the next sections emphasize the interactions between the two.

	No FE	FE
ρ	0.97	0.93
	(0.94, 1.00)	(0.90, 0.96)
σ	1.01	0.57
	(0.93, 1.09)	(0.53, 0.61)
$IQR_{0.75}$	1.22	0.69
	(1.12, 1.33)	(0.64, 0.73)
$IQR_{0.90}$	2.44	1.38
	(2.24, 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)

(a) India

	No FE	FE
ρ	0.71	0.50
	(0.67, 0.74)	(0.46, 0.55)
σ	1.78	1.18
	(1.69, 1.87)	(1.14, 1.23)
$IQR_{0.75}$	2.16	1.43
	(2.05, 2.27)	(1.38, 1.48)
$IQR_{0.90}$	4.31	2.86
	(4.10, 4.54)	(2.75, 2.97)
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, 1.13)

(b) Colombia

Notes: Panel (a): $N = 770$ households (India). Panel (b): $N = 2230$ households (Colombia). Results from the linear model (7), with and without fixed effects (FE). Year (survey round) dummies included in all specifications; month (interview) dummies also included for Colombia. 90% block bootstrap confidence intervals in parentheses.

TABLE 3. Results from the linear model

[15] That is, for $p^* = 0.5$ a standard deviation on the observed log odds gives the range $(-0.61, 0.61)$, which maps to probabilities as $(0.35, 0.65)$.

5.1.2 Colombia

In Table 3(b), we report estimation results for Colombia using the same two versions of the linear model — without and with fixed effects — as in panel (a) 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 1.78. 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 measurement error 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 1.78 to 1.18. Moreover, the variance of the individual fixed effect is much larger than in the Indian sample and, as in India, is much larger than the variance of the village component of the fixed effect (four times). Taken together, these results suggest higher risks and lower persistence in Colombian monthly earnings compared to Indian annual earnings, and greater unobserved heterogeneity in the Colombian data.

Controlling for observables in the linear model. For both India and Colombia, we extend the simple linear model considered so far by adding several observable variables, both on their own and interacted with income.

With this exercise we check whether the general properties of the linear model we estimate change substantially, or whether current income is a sufficient statistic for studying the dynamics of perceived future income. In particular, with a set of fully interacted dummies, we control for the number of income sources and the fraction of income coming from agriculture in India, and for the number of earners and the fraction of income coming from regular sources in Colombia.

The introduction of the additional controls interacted with current income does not make much difference to either the size of the uncertainty or to the dispersion of both household-level and village-level fixed effects. The introduction of these additional state variables seems to play a limited role in this model, both in India and Colombia. We report the results of this exercise in Appendix B. Given these results, we now focus on relaxing the linearity of the income process.

5.2 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:

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

which corresponds to equation (17) 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 $K = 3$ 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 (log current income enters in levels, but since $s_{jit} = r_{jit} - y_{i,t}$, quadratic 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 K and higher-order polynomials. While for $K > 3$ 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.

5.2.1 India

Table 4 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. That is, less poor households have less dispersion risk but more skewness risk than poorer ones.

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 higher-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 6, 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

	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)	

Notes: N = 770 households. Results for India from the flexible model with additive fixed effects in equation (16). Year (survey round) dummies included; reference group is the first year. 90% block bootstrap confidence intervals in parentheses.

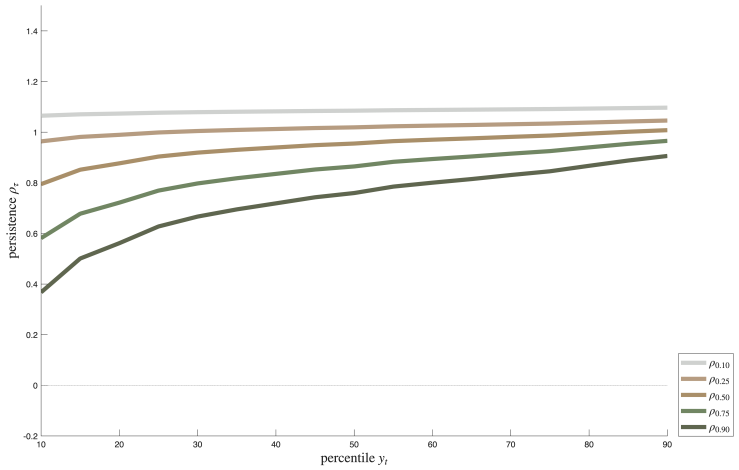
TABLE 4. Results from the flexible model with additive fixed effects, India

based on realized incomes because the reference populations in the two studies are very different; in our data 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 with 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 history and get the household (persistently) off the hook at a higher income level.[16] We find it quite interesting that this type of poverty-trap

[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



Notes: $N = 770$ households (two waves). The figure reports estimates of nonlinear persistence for India from the flexible model with additive fixed effects in equation (16). Year (survey round) dummies included; reference group is the first year. Pointwise confidence bands are shown in Figure C.2.

FIGURE 6. Nonlinear persistence (flexible model), India

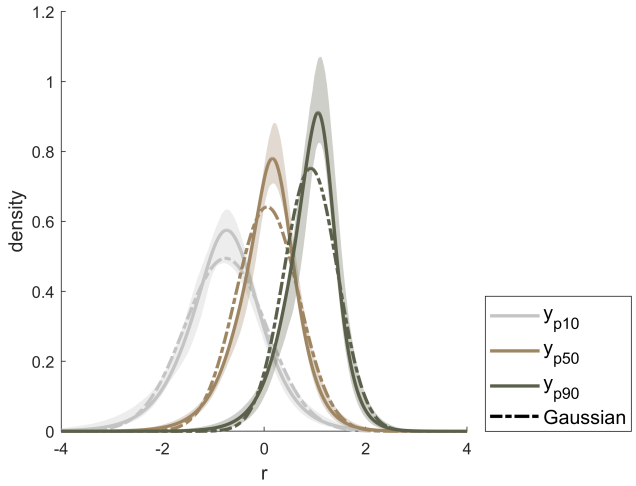
dynamic seem to be reflected in the subjective income expectations of poor households.

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 Supplemental 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.

Fitted density functions. Figure 7 shows the corresponding conditional probability density functions (*pdf*s) at the 10th, 50th, and 90th percentiles of current income for the baseline model, obtained by numerical differentiation of the estimated conditional cumulative distribution functions. Reference Gaussian distributions with matching mean and variance are included to assess departures from normality.

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 4, subjective predictive densities tend to be more symmetric and are noticeably more dispersed for the poor. Beyond non-normality, the figure also

an aspirations-based theory of poverty traps and references to the earlier theoretical literature.



Notes: $N = 770$ households. The figure plots fitted *pdf*s at the 10th, 50th, and 90th percentiles of current income for India. Results correspond to the flexible model with additive fixed effects in equation (16). Year (survey round) dummies included; reference group is the first year. Shaded areas indicate 90% pointwise confidence bands based on block bootstrap, and dotted lines show Gaussian densities with the same mean and variance as the corresponding empirical *pdf*s. The support is standardized future log income; see Supplemental Appendix B.4.

FIGURE 7. Probability density functions (flexible model), India

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, consistent with the patterns we find in Table 4.

5.2.2 Colombia

Table 5 presents the results for the nonlinear model with additive effects on the Colombian data. As in 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 with relatively good shocks for low-income households but not for high-income households (Table 5 and Figure 8). The impact of unobserved heterogeneity on persistence is also similar to that for India; see panel (b) in Figure C.2 in Supplemental Appendix C.

We also observe marked departures from normality in Colombia according to Figure 9, which

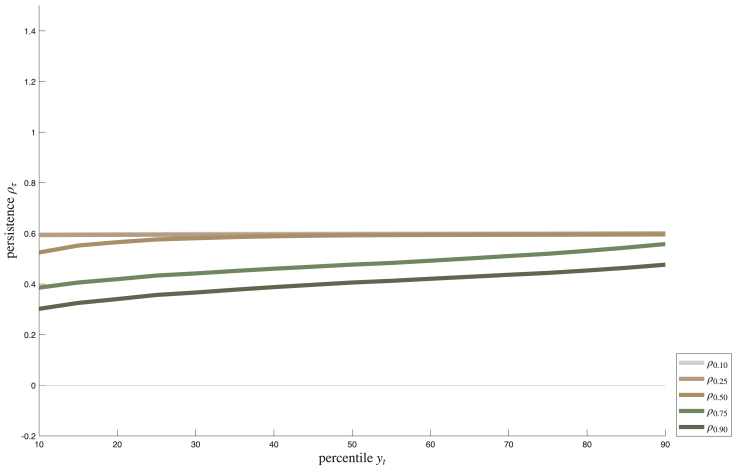
[17] Recall that, according to the chain rule in equation (15), nonlinear persistence can be obtained as the derivative of the *cdf* with respect to y relative to the derivative of the *cdf* with respect to r.

	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)	

Notes: N = 2230 households. Results for Colombia from the flexible model with additive fixed effects in equation (16). Year (survey round) and month (interview) dummies included; reference group is the first year and month. 90% block bootstrap confidence intervals in parentheses.

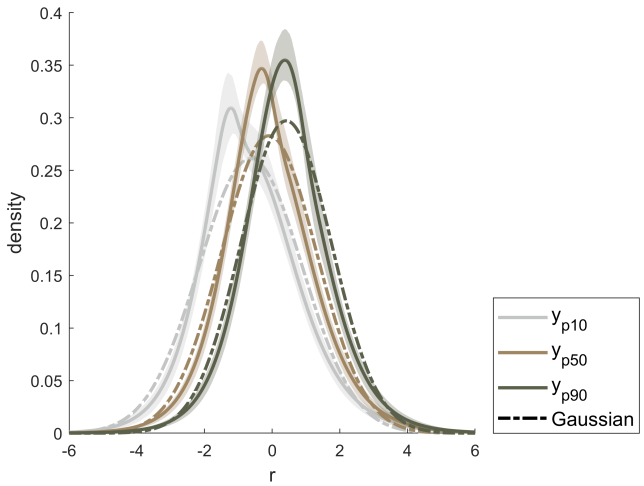
TABLE 5. Results from the flexible model with additive fixed effects, Colombia

shows fitted *pdf*s and reference Gaussian densities. The estimated densities are consistent with the pattern in Table 5 of decreasing dispersion risk as we move along the income gradient and feature prominent deviations from normality, more so for poorer households.



Notes: $N = 2230$ households. The figure reports estimates of nonlinear persistence for Colombia from the flexible model with additive fixed effects in equation (16). Year (survey round) and month (interview) dummies included; reference group is the first year and month. Pointwise confidence bands are shown in Figure C.2.

FIGURE 8. Nonlinear persistence (flexible model), Colombia



Notes: $N = 2230$ households. The figure plots fitted *pdf*s at the 10th, 50th, and 90th percentiles of current income for Colombia. Results correspond to the flexible model with additive fixed effects in equation (16). Year (survey round) and month (interview) dummies included; reference group is the first year and month. Shaded areas indicate 90% pointwise confidence bands based on block bootstrap, and dotted lines show Gaussian densities with the same mean and variance as the corresponding empirical *pdf*s. The support is standardized future log income; see Supplemental Appendix B.4.

FIGURE 9. Probability density functions (flexible model), Colombia

5.3 Generalizing heterogeneity patterns: interacted fixed effects

In this section, we discuss the estimation of the nonlinear model (17) with interacted fixed effects, in which $\beta_2^\dagger(s_{jit})$ is allowed to depend on s_{jit}. In these models, log-odds ratios can vary differentially with fixed effects and therefore allow for a greater distributional role of unobserved heterogeneity in accounting for nonlinearities.

As with the nonlinear model with additive effects, we report results for a parsimonious specification, with $K = 3$ for $\beta_0^\dagger(\cdot)$ and $\beta_1^\dagger(\cdot)$, $K = 2$ 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 — and is estimated via instrumental variables. We resort to the linear TSLS estimator introduced in Section A.2 (which does not impose the restrictions in equation (21)) using a full set of first-stage interaction terms.[18]

Tables 6 and 7 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 4. 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 7 tell a similar story relative to those in Table 5 for the nonlinear additive model.

State dependence versus unobserved heterogeneity. We have found that state dependence and unobserved heterogeneity compete as explanations for persistence, not only in linear models (as shown in Table 3), 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 quantify the relative contributions of each one to different features of a distributional income process estimated from subjective expectations data.

[18] Recall that we also include year indicators in all specifications, and that the TSLS estimator on the transformed equation (21) still requires us to account for all controls, even if the nonlinear restrictions are not imposed.

[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)	

Notes: $N = 770$ households. Results for India from the flexible model with multiplicative fixed effects in equation (17). Year (survey round) dummies included; reference group is the first year. 90% block bootstrap confidence intervals in parentheses.

TABLE 6. Results from the flexible model with multiplicative fixed effects, India

	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	1.91	1.62	1.52
	(1.19, 4.16)	(1.09, 2.89)	(1.10, 2.34)
$IQR_{0.90}$	3.85	3.57	3.48
	(2.45, 8.22)	(2.35, 6.59)	(2.37, 6.13)
$SK_{0.90}$	0.37	0.27	0.16
	(0.21, 0.56)	(0.14, 0.48)	(0.06, 0.34)
$\rho_{\tau 0.25}$	0.59	0.49	0.38
	(0.44, 0.69)	(0.23, 0.64)	(−0.10, 0.61)
$\rho_{\tau 0.50}$	0.50	0.58	0.49
	(−0.38, 0.67)	(0.40, 0.68)	(0.17, 0.65)
$\rho_{\tau 0.75}$	0.19	0.26	0.39
	(−1.11, 0.54)	(−0.70, 0.58)	(−0.39, 0.63)
σ^2_η		0.47	
		(0.41, 0.58)	
σ^2_η village		0.11	
		(0.12, 0.17)	
σ^2_ε		1.10	
		(1.05, 1.22)	

Notes: N = 2230 households. Results for Colombia from the flexible model with multiplicative fixed effects in equation (17). Year (survey round) and month (interview) dummies included; reference group is the first year and month. 90% block bootstrap confidence intervals in parentheses.

TABLE 7. Results from the flexible model with multiplicative fixed effects, Colombia

6 Conclusion

Households make intertemporal choices taking into account their own perceptions about future income. In this paper, we have developed an econometric framework for modeling income processes that allows us to obtain empirical measurements of subjective risk and subjective persistence directly from household-level subjective probabilistic expectations of future income. An important point we stress is that the availability of high-quality subjective expectations data makes it possible to identify flexible models of income dynamics under weaker assumptions and with shorter longitudinal data requirements than those needed to identify the same models using realized income alone. Moreover, the identified model captures income dynamics *as perceived by households*, and those perceptions are the ones that influence household decisions.

Using longitudinal survey data from India and Colombia, we find that linear income processes are soundly rejected, in line with the literature on flexible income processes using data on income realizations coupled with statistical models of the dynamics and rational expectations. 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 poor households anticipate lower persistence from large positive shocks, but this is not the case for richer households faced with 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 perceived 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. Taken together, our results suggest complex and heterogeneous patterns in the perceived transmission of income shocks to consumption, involving precautionary dispersion and skewness motives, which depend on the household's position in the income distribution.

Our work also suggests several avenues for future research. Despite considerable progress in the elicitation of subjective probability distributions in many different contexts, several important challenges remain, both in terms of questionnaire design and the econometric techniques that are used to analyze these data. Despite considering flexible models, elicitation errors remain large. Additional work is needed to understand the causes behind this, possibly jointly modeling them alongside measurement error in observed outcomes. Moreover, access to longer panel data on observed outcomes would allow a comparison between agents' perceived income processes and those inferred from realized income—an exercise that is not feasible with our two-wave panels.

On a different note, an interesting question is to what extent the features of income processes that we have documented are predictive of household choices. Along these lines, Attanasio, Sancibrián, and Ambrosio (2025) combine our estimates of subjective risk and persistence in

Colombia with additional household-level consumption data to estimate augmented Townsend (1994) regressions and study village-level risk-sharing arrangements. They find that deviations from perfect risk-sharing are smaller in villages where income processes are perceived to be less persistent or more volatile, in line with the predictions of models with self-enforceable, informal insurance contracts.

Appendix

A 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.

A.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 probabilities p_{jit}, which are noisy measures of $F_{it}\left(r_{jit}\right) = P\left(y_{i,t+1} < r_{jit} | I_{it}\right)$, also measure $F_{it}^{\Delta}\left(s_{jit}\right) = P\left(\Delta y_{i,t+1} < s_{jit} | I_{it}\right)$ 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 nonparametric 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 with modeling. Furthermore, while 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, they require translations of the index in the *cdf* of $y_{i,t+1}$.[20]

[20] 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 by additional nonlinear variation in the tails.

In the linear case, the model to be estimated remains essentially unchanged by targeting log income changes instead of log levels.[21] However, for nonlinear specifications, the actual flexible model to be estimated will be different:

$$\ell_{jit} = \beta_0^\dagger(s_{jit}) + \beta_1^\dagger(s_{jit})\psi(y_{it}) + \beta_2^\dagger(s_{jit})\eta_i + \varepsilon_{jit}. \tag{17}$$

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 is based on equation (17), taking the components of $\psi(\cdot)$ in a polynomial basis of functions. We specify $\psi(\cdot)$ 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 $K \geq 2$ knots requires estimating K parameters. Further details are provided in Supplemental 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 $\Lambda(\cdot)$.

A.2 Estimation

In specifications with $\beta_2^\dagger(s_{jit}) = 1$, this remains a static linear panel data model whose parameters 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. Regardless of whether the underlying probabilities satisfy the monotonicity property of *cdf*s, we estimate the curve g pointwise for every point in the (discretized) support of r; the resulting estimate is therefore not guaranteed to be strictly increasing. We follow the method proposed by Chernozhukov, Fernández-Val, and Galichon (2010), which consists of 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.

[21] The linear reparameterized equation is

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

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

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 is due 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-variables (IV) estimator of the kind that we employ in obtaining our empirical results, and defer a more general discussion of the estimation of models with interacted effects to Supplemental Appendix B.

A simple IV estimator. Let us consider the model

$$\ell_{jit} = \beta_0 r_{jit} + \beta_1 y_{it} + \left(1 + \beta_2 r_{jit}\right)\eta_i + \varepsilon_{jit}, \tag{18}$$

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 (18):

$$\tilde{\ell}_{jit} = \beta_0 \tilde{r}_{jit} + \beta_1 \tilde{y}_{it} + \beta_2 \tilde{r}_{jit}\eta_i + \tilde{\varepsilon}_{jit}, \tag{19}$$

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

$$r_{jit}\bar{\ell}_i - \bar{r}_i\ell_{jit} = \beta_1\left(r_{jit}\bar{y}_i - \bar{r}_i y_{it}\right) + \tilde{r}_{jit}\eta_i + \left(r_{jit}\bar{\varepsilon}_i - \bar{r}_i\varepsilon_{jit}\right) \tag{20}$$

to substitute out $\tilde{r}_{jit}\eta_i$ in (19) and obtain

$$\tilde{\ell}_{jit} = \beta_0\tilde{r}_{jit} + \beta_1\tilde{y}_{it} + \gamma\left(r_{jit}\bar{y}_i - \bar{r}_i y_{it}\right) + \beta_2\left(r_{jit}\bar{\ell}_i - \bar{r}_i\ell_{jit}\right) + \xi_{jit}, \tag{21}$$

where $\xi_{jit} = (1 + \beta_2\bar{r}_i)\,\varepsilon_{jit} - (1 + \beta_2 r_{jit})\bar{\varepsilon}_i$ 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 (21). 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 averages of w_i over t and j 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 (21) by two-stage least squares (TSLS). This follows from the fact that endogeneity is driven by the presence of ε_{jit} (and $\bar{\varepsilon}_i$) in equation (20), but $E\left[\varepsilon_{jit}\middle|w_i\right] = 0$ and thus any predictor of η_i using w_i (conditional on the included regressors) is a potentially valid instrument. The restriction $\gamma = -\beta_1\beta_2$ is not required for identification and can be ignored. We do so in our implementation to preserve the linearity of the approach.

B Additional results: 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. Along these lines, additional flexibility can be added by including relevant time-varying household characteristics x_{it} in the conditioning set, which extends the baseline linear model to

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

This simply reproduces equation (8) in the main text, and allows for differential subjective persistence along these characteristics. We report below results for linear models when including information on the type and number of sources of income and the type of shocks experienced by households.

B.1 India

In Table 8, we introduce indicators of the type and number of sources of household income, whereas in Table 9 we consider interactions of current income with different types of negative income shocks experienced by households.

Remarkably, the introduction of these additional variables does not greatly affect the variability of income innovations, the extent of persistent unobserved heterogeneity, or the variance of elici-

tation errors. In sum, although some coefficients are marginally significant, these additional state variables do not seem to have a large effect on the estimated risk and persistence of the income process.

ρ	≤2 sources	3 sources	4+ sources
0% farm	0.87	0.90	0.83
	(0.83, 0.91)	(0.84, 0.95)	(0.76, 0.89)
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, 0.95)
σ		0.55	
		(0.51, 0.58)	
$IQR_{0.90}$		1.33	
		(1.23, 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)	

Notes: $N = 770$ households. Results for India from the linear model (7), augmented with household-level characteristics following (8). "Farm" denotes the share of current income from farming-related activities; see Table 1 for variable definitions. Year (survey round) dummies included. 90% block bootstrap confidence intervals in parentheses.

TABLE 8. Results from the linear model with additional household-level characteristics (share of income from farming and income sources), India

B.2 Colombia

Table 11 considers interactions of current income with the number of earners and the percentage of income accrued from regular work, whereas Table 10 considers interactions with the number of earners and the type of negative shocks experienced by households.

These results seem to suggest a higher level of persistence for households composed of three or more working members (which might point to a case of income diversification, as would be natural in a context where predictions are one month ahead), but overall the introduction of these additional state variables seem to play a limited role in this model, similar to our discussion for the Indian data.

ρ	≤2 sources	3 sources	4+ sources
No shock	0.87	0.91	0.83
	(0.79, 0.95)	(0.83, 0.99)	(0.73, 0.92)
Health	0.92	0.97	0.89
	(0.86, 0.97)	(0.90, 1.03)	(0.81, 0.97)
Agricultural	0.90	0.97	0.87
	(0.86, 0.95)	(0.92, 1.02)	(0.81, 0.94)
Other	0.99	1.04	0.97
	(0.87, 1.10)	(0.92, 1.15)	(0.83, 1.08)
σ		0.55	
		(0.51, 0.58)	
$IQR_{0.90}$		1.34	
		(1.23, 1.41)	
σ^2_η		0.25	
		(0.22, 0.32)	
σ^2_η village		0.15	
		(0.14, 0.20)	
σ^2_ε		1.13	
		(1.07, 1.16)	

Notes: N = 770 households. Results for India from the linear model in (7), augmented with household-level characteristics following (8). Variable definitions are provided in Table 1. Year (survey round) dummies included. 90% block bootstrap confidence intervals in parentheses.

TABLE 9. Results from the linear model with additional household-level characteristics (shocks and income sources), India

ρ	1 earner	2 earners	3+ earners
No shock	0.54	0.55	0.58
	(0.47, 0.61)	(0.47, 0.63)	(0.48, 0.68)
Health	0.64	0.65	0.67
	(0.51, 0.76)	(0.50, 0.77)	(0.52, 0.81)
Other	0.44	0.46	0.48
	(0.33, 0.56)	(0.33, 0.57)	(0.34, 0.61)
σ		1.17	
		(1.12, 1.21)	
$IQR_{0.90}$		2.84	
		(2.72, 2.94)	
σ^2_η		0.48	
		(0.45, 0.53)	
σ^2_η village		0.11	
		(0.12, 0.16)	
σ^2_ε		1.09	
		(1.05, 1.12)	

Notes: N = 2230 households. Results for Colombia from the linear model in (7), augmented with household-level characteristics following (8). Variable definitions are provided in Table 2. Year (survey round) and month (interview) dummies included. 90% block bootstrap confidence intervals in parentheses.

TABLE 10. Results from the linear model with additional household-level characteristics (shocks and number of earners), Colombia

ρ	1 earner	2 earners	3+ earners
0% regular	0.34	0.36	0.48
	(0.20, 0.47)	(0.22, 0.49)	(0.32, 0.63)
75% regular	0.51	0.52	0.61
	(0.44, 0.57)	(0.44, 0.59)	(0.50, 0.72)
100% regular	0.56	0.57	0.65
	(0.50, 0.63)	(0.48, 0.65)	(0.54, 0.76)
σ		1.17	
		(1.12, 1.21)	
$IQR_{0.90}$		2.83	
		(2.72, 2.93)	
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)	

Notes: $N = 2230$ households. Results for Colombia from the linear model (7), augmented with household-level characteristics following (8). See Table 2 for variable definitions. Year (survey round) and month (interview) dummies included. 90% block bootstrap confidence intervals in parentheses.

TABLE 11. Results from the linear model with additional household-level characteristics (share of income from regular sources and number of earners), Colombia

References

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

Altonji, J., D. Hynsjö, 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*, 85, 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.

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.

Attanasio, O., A. Kovacs, and K. Molnar (2020): "Euler Equations, Subjective Expectations and Income Shocks." *Economica*, 87, 406–441.

Attanasio, O., V. Sancibrián, and F. Ambrosio (2025): "New measures for richer theories: some thoughts and an example." Working paper.

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

Bachmann, R., G. Topa, and W. van der Klaauw (2023): *Handbook of Economic Expectations*, Elsevier, 1st ed.

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.

Caplin, A., V. Gregory, E. Lee, S. Leth-Petersen, and J. Sæverud (2024): "Subjective Earnings Risk." Working paper.

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

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.

De Nardi, M., G. Fella, 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 R. Bachmann, G. Topa, and W. van der Klaauw, 261–291.

Delavande, A., X. Giné, and D. McKenzie (2011a): "Eliciting probabilistic expectations with visual aids in developing countries: How sensitive are answers to variations in elicitation design?" *Journal of Applied Econometrics*, 26, 479–497.

——— (2011b): "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.

D'Haultfoeuille, X., C. Gaillac, and A. Maurel (2021): "Rationalizing Rational Expectations: Characterizations and Tests." *Quantitative Economics*, 12, 817–842.

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.

Fuster, A. and B. Zafar (2023): "Survey Experiments on Economic Expectations." in *Handbook of Economic Expectations*, ed. by R. Bachmann, G. Topa, and W. van der Klaauw, Elsevier, 4, 107–130.

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.

Jappelli, T. and L. Pistaferri (2000): "Using subjective income expectations to test for excess sensitivity of consumption to predicted income growth," *European Economic Review*, 44, 337–358.

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

Kosar, G. and W. van der Klaauw (2025): "Workers' Perceptions of Earnings Growth and Employment Risk." *Journal of Labor Economics*, 43, 83–121.

Kosar, G. and C. O'Dea (2023): "Expectations Data in Structural Micro Models." in *Handbook of Economic Expectations*, ed. by R. Bachmann, G. Topa, and W. van der Klaauw, Elsevier, 21, 647–675.

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.

Pistaferri, L. (2003): "Anticipated and Unanticipated Wage Changes, Wage Risk, and Intertemporal Labor Supply," *Journal of Labor Economics*, 21, 729–754.

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

Townsend, R. M. (1994): "Risk and Insurance in Village India," *Econometrica*, 62, 539–591.