Production function estimation using subjective expectations data

Agnes Norris Keiller* Áureo de Paula† John Van Reenen‡

April 15, 2026

Abstract

Standard methods for estimating production functions in the Olley and Pakes (1996) tradition require assumptions on input choices. We introduce a new method that exploits (increasingly available) data on a firm's expectations of its future output and inputs that allows us to obtain consistent production function parameter estimates while relaxing these input demand assumptions. In contrast to dynamic panel methods, our proposed estimator can be implemented on very short panels (including a single cross-section), and Monte Carlo simulations show it outperforms alternative estimators when firms' material input choices are subject to optimization error. Implementing a range of production function estimators on UK data, we find our proposed estimator yields results that are either similar to or more credible than commonly-used alternatives. These differences are larger in industries where material inputs appear harder to optimize. We show that TFP implied by our proposed estimator is more strongly associated with future growth than existing methods, suggesting that failing to adequately account for input endogeneity may underestimate the degree of dynamic reallocation in the economy.

JEL classification: L11, L23, C23, C21, O31.
Keywords: Expectations, Production Functions, Productivity, Input demand

Acknowledgements: We would like to thank Orazio Attanasio, Nick Bloom, Richard Blundell, Moshe Buchinsky, Michele Fioretti, Bruce Hansen, Bo Honoré, Joel Horowitz, Isabela Manelici, Chuck Manski, Ariel Pakes, Joris Pinkse, Maarten de Ridder, Jean-Marc Robin and participants at several institutional seminars and conferences for comments. Generous funding has come from the ESRC and ERC through the Programme for Innovation and Diffusion (Grant Ref: ES/V009478/1) and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Ref: EP/X02931X/1) respectively. This work was undertaken in the Office for National Statistics Secure Research Service using data from ONS and other owners and does not imply the endorsement of the ONS or other data owners.

*London School of Economics
†University College London, Institute for Fiscal Studies and CeMMAP
‡London School of Economics and MIT

1 Introduction

The 'production function' – a representation of the process by which inputs are turned into outputs – has long been an object of great economic interest (Cobb and Douglas 1928; Griliches and Mairesse 1999). Production functions are critical to examining a wide range of topics including technological change, productivity dispersion, firm markups and the impact of policy. Research on these topics has gained greater salience in recent years, in part due to the productivity growth slowdown, particularly since the 2008-9 Global Financial Crisis. Before one can analyze such topics, however, it is necessary to consistently estimate a production function, which has proven no easy task.[1]

Econometric research on production function estimation has had a renaissance in recent years.[2] Estimation is complicated by a number of long-known issues, most notably the endogeneity of inputs: because a firm's productivity is unobservable and likely correlated with input choices, straightforward estimation methods such as OLS regression will be biased (Marschak and Andrews 1944; Zellner et al. 1966).

Standard methods to deal with these problems included controlling for firm fixed effects (Mundlak 1961) by differencing and instrumenting with lagged input values (Anderson and Hsiao 1981), for example. Such approaches generally find implausibly low estimates for relevant parameters, especially on the output elasticity with respect to capital.[3] Blundell and Bond (2000) consider an alternative approach in this dynamic panel data literature by including lagged differences as instruments for the levels of factor inputs. This nonetheless requires conditioning on at least three consecutive time series observations on a firm, which in many empirical settings loses a considerable subset of data. Moreover, it relies on exact parametric specification of the productivity process and requires a strong stationarity assumption making the method potentially inappropriate for younger and fast-growing firms.

1. Production functions have also found insightful applications beyond firm-level analysis to understand, for example, the impact of various types of inputs on child development (Todd and Wolpin 2003; Cunha and Heckman 2008). The measurement challenges and context there are somewhat distinct and the techniques used in estimation differ from those used in the analysis of firm production. In this context, Attanasio et al. (2024) recently use subjective beliefs to examine parents' perception of the skills formation production function.

2. For surveys see De Loecker and Syverson (2021) or Ackerberg et al. (2007). Recent contributions include Ackerberg et al. (2015), Collard-Wexler and De Loecker (2021), De Loecker (2011), De Loecker et al. (2016), Doraszelski and Jaumandreu (2013, 2018), Gandhi et al. (2020), Orr (2022), De Roux et al. (2021) and Valmari (2023).

3. This has generally been thought to be because of the high persistence of the capital stock. Differencing removes all cross sectional information on capital, and much of the remaining time series variation may be measurement error. Moreover, lags will be poor predictors of the change in the capital stock, if the level of capital is close to a random walk.

The drawbacks of these dynamic panel data methods have contributed to the popularity of an alternative suite of 'proxy variable' production function estimators that use a non-parametric function of various observables to control for unobserved productivity. The pioneers of this method were Olley and Pakes (1996) (OP), who control for productivity using a flexible function of investment and capital that represents the inverse of firms' investment policy. The reasoning behind their approach is that if firms' investment policy function can be written as an invertible function of pre-determined capital and the persistent component of unobserved productivity, then the latter can be proxied with a flexible function of capital and investment. Levinsohn and Petrin (2003) (LP) instead propose using the inverse of firms' material input demand as a proxy for productivity to address selection concerns implicit in the OP approach due to potentially high prevalence of zero investment among firms.

Noting that a production function's labor input parameter is unidentified using the LP methodology under plausible assumptions, Ackerberg et al. (2015) (ACF) outline a refinement on timing assumptions and proxy variable arguments to address this. Unlike LP, who rely on a material input demand function conditioned solely on capital, ACF propose controlling for unobserved productivity by inverting a material input demand function conditioned on labor as well as capital and then recovering both input elasticities in a second estimation stage. Gandhi et al. (2020) (GNR) show that this suite of estimators may fail to identify the parameters of a 'gross output' production functions (i.e. one that includes materials as an input), and propose an alternative estimation strategy based on the implications of price-taking firms' optimality conditions. Bond and Söderbom (2005) provide another estimator, showing that in the presence of adjustment costs on all inputs the parameters of a Cobb-Douglas production function can be recovered by using lagged levels of inputs as instruments. This approach incorporates aspects of both the dynamic panel literature – in using lags as instruments and specifying the productivity process – and the proxy variable approach – by relying on the implications of optimal firm input decisions to yield identification. However, simulation results show that their proposed method is sensitive to the form and magnitude of adjustment costs which may explain why it has not been widely deployed. More in-depth reviews of alternative production function estimation strategies are provided in Ackerberg et al. (2007) and De Loecker and Syverson (2021).

Despite their differences, proxy methods such as OP, LP and ACF all rely on the existence of a strictly monotonic relationship between a firm's (conditional) input demand and productivity – an assumption justified with recourse to models of firms' decisions that yield optimal policies satisfying monotonicity. The performance of these estimation

methods is therefore threatened by any unobserved factor that violates the required relationship between productivity and the input used to generate its proxy, such as input adjustment costs or prices that vary across firms and optimization errors, and may suffer if relevant variables (e.g. factor prices) are omitted or unavailable.[4]

This paper contributes to the literature on production function estimation by showing how data on firms' perceptions of its future output and inputs can be used to recover consistent production function parameter estimators while relaxing assumptions on firms' input demand policies. We leverage information in recent surveys that collect detailed information on firms' perceived *probabilistic distribution* of output (e.g. revenues) and inputs (e.g., employment, intermediates and capital expenditure) in the future. The intuition underpinning our approach is that firms' expectations regarding future inputs and output contain information about their expected future productivity which, in turn, contains information about their current productivity. Unlike dynamic panel estimators, which require parametric specification of the productivity process, we require the relatively common assumption that persistent productivity follows a first-order Markov process – an assumption imposed as well in OP, LP, ACF and GNR. Combined with assumptions that persistent productivity is unidimensional and that there is a monotonic relationship between current and expected future productivity, which are also imposed by standard proxy variable methods, firms' expectations can be used to control for unobserved productivity and thereby recover consistent parameter estimates.[5] Our proposed method is therefore similar to OP/LP/ACF as it requires a monotonic relationship between productivity and observables, but different as it leverages data on firms' expectations rather than optimal input quantities. It is therefore robust to a range of factors – such as unobserved firm-specific input prices and optimization error – that would undermine alternative estimators by breaking the one-to-one link between input demands and productivity. Unlike the dynamic-panel and control function estimators, our proposed method can be implemented in a single cross-section of data. This enables production function estimation in contexts where panel data would be infeasible to collect for either

4. Gandhi et al. (2020) by contrast, explicitly require firms' flexible input demands to be optimal and is therefore compromised by any factor causing deviations from optimality.

5. The requirement that persistent productivity be unidimensional confines us to a setting where productivity shocks are Hicks-neutral. While this is conventional in the literature, two notable exceptions that accommodate factor-augmenting technology shocks are Doraszelski and Jaumandreu (2018) and Demirer (2022). Doraszelski and Jaumandreu (2018) relaxes the assumption of unidimensional productivity by trying to leverage data on firm-level input prices while Demirer (2022) does so by imposing assumptions on firms' input demands. While both these papers are valuable contributions and argue the importance of factor-augmenting productivity shocks, identifying two sources of technical change is more challenging (e.g., average wages reflect labor composition changes as well as pure labor cost shocks).

cost or logistical reasons. We also note that the moments we exploit here can be seen as complementary to those already in use in the traditional methods whenever data is available to implement both, thus lending more precision or over-identifying restrictions to existing methods.

Monte Carlo simulations show that our proposed estimation method recovers precise estimates of production function parameters under a range of data generating processes. Notably, it retains consistency when firms' input decisions are subject to optimization error whereas other approaches generally do not. While our proposed estimator is undermined if firms' expectations of next-period productivity are biased, we show our basic estimation algorithm can be extended to retain consistency if this bias is a function of a time-invariant attribute and if one possesses panel data on realizations and expectations of revenue and employment.

To test the empirical performance of our method, we leverage the UK's Management and Expectations Survey (Office For National Statistics 2022). The MES records information on firms' inputs, output and their one-year-ahead expectations of these quantities between 2016 and 2020. We focus on three industries – electronics manufacturing, wholesale and retail and restaurants – and estimate industry-specific production functions using a range of methods. The estimates recovered using our proposed method are broadly similar for the electronics and retail production functions but differ in non-negligible yet plausible ways for the restaurant sector. To rationalize these results, we show material inputs are subject to particularly large within-year revisions among the restaurant sector, which suggests optimization is particularly hard for these firms and hence the LP and ACF monotonicity assumption less likely to hold. We use the alternative production function estimates to recover estimates of total factor productivity (TFP), and compare static and dynamic moments of its distribution across methods. We relate TFP estimates to firm performance and find estimates obtained using our proposed estimator are more positively associated with future employment growth than alternative estimators, particularly over a four-year horizon.

Combined with the Monte Carlo evidence, our empirical application demonstrates the utility of expectations data in the context of production function estimation and thereby contributes to a more general literature documenting the value of expectations data. Starting in the 1990s much of this literature's initial focus was on income dynamics with, for example, the seminal work by Dominitz and Manski (1997), who demonstrate how surveys can be used to elicit subjective income expectations, and Pistaferri (2001), who shows the econometric benefits of such additional information as a means to separately identify permanent and transitory shocks to income. Manski (2004) summarised these

early advances and argued that data on expectations could be useful both as a means to relax and validate assumptions within various economics models. Of the subsequent work that has examined the value of subjective expectations data in a wide range of contexts, our work is related to Gennaioli et al. (2016), in that it demonstrates insights that can be gained from firms' expectations rather than those of individuals' (in particular, Bloom et al. (2025)). Our work is also related to recent and ongoing work by Arellano et al. (2024), who return to the literature's early focus on subjective income expectations and show how such data can be used to estimate income processes in a flexible manner that relaxes commonly-imposed parametric assumptions. Similar to their work, we document that the additional information contained within data on subjective expectations allows one to relax particular assumptions that underpin conventional production function estimators and thus implicitly allows for more flexible models of firm behavior.

The remainder of the paper is as follows. In section 2 we show how expectations data identify production function parameters, describe our proposed estimation methodology and compare it to other standard methods. Section 3 outlines the Monte Carlo setup we use to compare alternative estimators and discusses the results across various data generating processes. Section 4 describes the data we use in our empirical application, the results of which are described in section 5. Section 6 concludes. In online Appendices we provide a discussion of expectation formation processes (D), a convergence proof of our estimator (B), NPR implementation details (C), extensions to the estimator, in particular to biased expectations (D), Monte Carlo details (E) and Data and Supplementary results (F).

2 Production function estimation using firms' expectations

Consider a general production function of the following form

$$y_{it} = f(k_{it}, l_{it}; \beta) + \omega_{it} + \epsilon_{it}, \tag{1}$$

where subscript i denotes firm and subscript t denotes time. Lower case letters denote logs, so y is the log of output, k is the log of capital, l is the log of labor and $f(\cdot; \beta)$ is some general function of the two with parameters β, which captures the process by which they are combined during production.[6] The variables ω and ϵ are unobserved by

6. For the general exposition of this subsection, output may either be value added (i.e. net output) or turnover (the British term for sales revenue or gross output, which we use throughout this paper). In the latter case, the omission of materials as an input can be justified under a Leontief model of production in which labor and capital are combined in a fixed proportion with materials (Ackerberg et al. 2015). In

the econometrician. The variable ω represents idiosyncratic productivity that is known by the firm at the time period-t input and investment decisions are made. In contrast, ϵ are unanticipated mean-zero disturbances representing productivity shocks, such as extreme weather events or machine failures, which only become observable to the firm *after* its period-t decisions have been made. Alternatively, ϵ can represent mean-zero measurement error, which does not affect the firm but poses problems to the econometrician (Griliches and Mairesse 1999).

Capital evolves according to

$$K_{it} = (1 - \delta)K_{it-1} + I_{it-1},$$

where δ is the depreciation rate and I_{it-1} is investment.[7] Unobserved productivity ω follows a Markov process

$$\omega_{it} = \mathbb{E}[\omega_{it}|\Omega_{it-1}] + \xi_{it} = \mathbb{E}[\omega_{it}|\omega_{it-1}] + \xi_{it} = g(\omega_{it-1}) + \xi_{it},$$

where $\mathbb{E}[\xi_{it}|\Omega_{it-1}] = 0$ and Ω_{it-1} represents the firm's information set at $t-1$. The information includes $k_{it-1}, l_{it-1}, \omega_{it-1}, I_{it-1}$ (and thus k_{it}) but also additional variables such as input and output prices and demand factors.

2.1 Identification

The method we propose relies on a simple intuition, and there are two parts to this. First, data on a firm's expectations about its own output and inputs one period ahead $(t+1)$ are informative about expected productivity one period ahead $(t+1)$. If productivity follows a Markov process, those data thus also carry information about the firm's own productivity in the current period (t). Second, once period t productivity is controlled for using expectations for $t+1$, current inputs are orthogonal to the remaining error, which correspond to random shocks to productivity not observed by the firm before making input choices. The expectations data are useful to proxy for unobserved productivity in the spirit of other production function proxy methods. The first element is the more fundamental contribution as we bring new data to the estimation of production functions, rather than the traditional approach of just using panel data on stocks and flows of inputs

practice, the distinction will influence how data on firm expectations' are treated, which we return to in section 4.

7. Whereas we follow the literature in assuming "time-to-build", since the MES also collects information on capital expenditures, it is conceivable that this assumption may also be relaxed along the lines of our derivations below.

and output.

Suppose firms form expectations about their period $t+1$ production and inputs at the end of period t conditional on their information set $\Omega_{it} \supset \{k_{it}, l_{it}, I_{it}, \omega_{it}, k_{it+1}\}$. As is usually the case, we assume that firms know their production function $f(\cdot;\beta)$ and its parameter values, so they both belong to the information set. We will discuss imperfect knowledge in subsection 2.5.

Following the notation in Pesaran and Weale 2006, we append the subscripts i and t to highlight that the relevant variables are obtained with respect to the subjective probabilities reported by decision makers in firm i at period t (i.e., $\mathbb{E}_{it}[\mathbf{x}_{it+1}|\Omega_{it}] = \int \mathbf{x} h_{it}(\mathbf{x}|\Omega_{it})d\mathbf{x}$, where $\mathbf{x}$ is a vector and $h_{it}(\cdot)$ represents firm i's distribution for it in period t given their information Ω_{it}). As long as they satisfy the identification conditions in Theorem 1 and Corollary 1 (see below), we do not take a stand on the expectation formation process for output or inputs and treat those as data. One may nonetheless envision different scenarios on the expectation process for productivity and we discuss different expectation formation protocols in Appendix A: rational expectations, static expectations, "return to normality" and adaptive expectations.

For simplicity, in what follows we assume firms' expectations to be rational (i.e., $\mathbb{E}_{it}[\omega_{it}|\Omega_{it-1}] = \mathbb{E}[\omega_{it}|\Omega_{it-1}]$), which implies

$$\begin{aligned} \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] &= \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) + \mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] + \mathbb{E}_{it}[\epsilon_{it+1}|\Omega_{it}] \\ &= \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) + g(\omega_{it}), \end{aligned} \tag{2}$$

where $F_{it}(l_{it+1})$ represents firm i's subjective probability distribution over their next-period labor input given its information Ω_{it} and the second equality follows from the assumptions that $\mathbb{E}_{it}[\epsilon_{it}|\Omega_{it-1}]$ and $\mathbb{E}_{it}[\xi_{it}|\Omega_{it-1}]$ are equal to zero. The former relies on ϵ_{it} being unanticipated by definition and thus orthogonal to the information set Ω_{it} as discussed earlier. The latter follows from rational expectations about ω_{it} and thus $\mathbb{E}_{it}[\xi_{it}|\Omega_{it-1}] = \mathbb{E}[\xi_{it}|\Omega_{it-1}] = 0$.

Rearranging equation (2) for $g(\omega_{it})$ obtains

$$g(\omega_{it}) = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}). \tag{3}$$

Like other proxy variable approaches, we now require a monotonicity assumption. In our case this assumption is that the right hand side of equation (3) is strictly increasing in ω_{it}. Or, in words: given a firm's current (persistent) productivity there is a single level

of productivity they expect next period and this single level can be uniquely inferred from their expectations about next-period output, labor and the deterministic level of next-period capital. A related assumption – i.e., that the conditional distribution of ω_{it+1} given ω_{it} be increasing in the latter – appears in ACF (see their Assumption 2). It is used there and in OP to guarantee their invertibility condition (see Pakes (1994)), but it is not needed in their estimation protocol and typically it is not imposed.

Under the strict monotonicity assumption, ω_{it} can be recovered as

$$\omega_{it} = g^{-1}\left(\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1})\right). \tag{4}$$

If output (y_{it}), inputs (k_{it}, k_{it+1} and l_{it}), and beliefs ($\mathbb{E}_{it}[y_{it+1}|\Omega_{it}]$ and $F_{it}(\cdot)$) are observable, we can combine equations (1) and (4) to obtain a moment condition that can be used to recover the parameters of interest:

$$\begin{aligned}
&\mathbb{E}[\epsilon_{it}|\Omega_{it}] = \mathbb{E}[y_{it} - f(k_{it}, l_{it}; \beta) - \omega_{it}|\Omega_{it}] \\
&= \mathbb{E}\left[y_{it} - f(k_{it}, l_{it}; \beta) - \Psi\left(\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1})\right) \Big| \Omega_{it}\right] \\
&= 0,
\end{aligned} \tag{5}$$

where Ψ is some non-parametric function representing $g^{-1}(\cdot)$.

Consider for example the case of a Cobb-Douglas production function where $\beta = (\alpha, \beta_k, \beta_l)$ and an AR(1) process for ω with auto-regressive parameter ρ (i.e., $g(\omega) = \rho\omega$), then equation (5) becomes

$$\mathbb{E}\left[y_{it}|\Omega_{it}\right] - \frac{\alpha(1+\rho)}{\rho} - \beta_k k_{it} - \beta_l l_{it} - \frac{1}{\rho}\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] + \frac{\beta_k}{\rho} k_{it+1} + \frac{\beta_l}{\rho}\mathbb{E}_{it}[l_{it+1}|\Omega_{it}] = 0, \tag{6}$$

as long as $\rho \neq 0$. The model would then identify $\theta = (\beta, g)$ (where $g(\omega) = \rho\omega$) if, for example, $\mathbb{E}[x_{it}x_{it}^{\top}]$, where $x_{it} = (1, k_{it}, l_{it}, \mathbb{E}_{it}[y_{it+1}|\Omega_{it}], k_{it+1}, \mathbb{E}_{it}[l_{it+1}|\Omega_{it}])$, has full rank.[8] This example highlights that identification of θ by equation (5) will depend on the specifications of the production function $f(\cdot; \beta)$, the Markov process encoded in $g(\cdot)$ and on the degree of variation observed in the data.

For more general specifications, one can establish that:

Theorem 1. *Assume that $g(\cdot)$ is strictly monotonic and $\mathbb{E}[\epsilon_{it}|\Omega_{it}] = 0$. Let $x_{it} = (k_{it}, l_{it})$ and $z_{it} = (\mathbb{E}_{it}[y_{it+1}|\Omega_{it}], k_{it+1}, F_{it}(\cdot))$ and denote by $\theta_0 = (\beta_0, g_0)$ the data generating*

8. This is sufficient, but not necessary since there are four parameters in this specification.

parameters. Then, if

$$f(x_{it};\beta_0) - \mathbb{E}\big[f(x_{it};\beta_0)|z_{it}\big] \neq f(x_{it};\beta) - \mathbb{E}\big[f(x_{it};\beta)|z_{it}\big] \tag{7}$$

with positive probability for any $\beta \neq \beta_0$, the parameter vector $\theta_0 = (\beta_0, g_0)$ is identified.

Proof. Since $y_{it} = f(x_{it};\beta_0) + h_0(z_{it}) + \epsilon_{it}$, where

$$h_0(z_{it}) = \Psi_0\left(\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1};\beta_0) dF_{it}(l_{it+1})\right).$$

Taking expectations conditional on z_{it} on both sides and subtracting, one obtains that

$$\underbrace{y_{it} - \mathbb{E}(y_{it}|z_{it}))}_{\equiv w_{it}} = \underbrace{f(x_{it};\beta_0) - \mathbb{E}(f(x_{it};\beta_0)|z_{it})}_{\equiv m(x_{it},z_{it};\beta_0)} + \epsilon_{it},$$

where we use the fact that $h_0(z_{it}) = \mathbb{E}[h_0(z_{it})|z_{it}]$. Since $\mathbb{E}[\epsilon_{it}|\Omega_{it}] = 0$ and $\{x_{it}, z_{it}\} \subset \Omega_{it}$, we have that $\mathbb{E}[\epsilon_{it}|x_{it}, z_{it}] = 0$ and $m(x_{it}, z_{it};\beta_0) = \mathbb{E}(w_{it}|x_{it}, z_{it})$. It thus uniquely solves $\min_{\tilde{m}(\cdot)} \mathbb{E}[(w_{it} - \tilde{m}(x_{it}, z_{it}))^2]$ as long as condition (7) is satisfied with positive probability, which implies that β_0 is identified.

The function $\Psi_0(\cdot) \equiv g_0^{-1}(\cdot)$ is then identified since

$$\underbrace{y_{it} - f(x_{it};\beta_0)}_{\equiv \tilde{y}_{it}} = \Psi_0\left(\underbrace{\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1};\beta_0) dF_{it}(l_{it+1})}_{\equiv \tilde{z}_{it}}\right) + \epsilon_{it}.$$

Since $\tilde{z}_{it} \subset \Omega_{it}$, thus have that $\mathbb{E}[\epsilon_{it}|\tilde{z}_{it}] = 0$ and $\Psi_0(\tilde{z}_{it}) = \mathbb{E}(\tilde{y}_{it}|\tilde{z}_{it})$ and $g_0(\cdot) = \Psi_0^{-1}(\cdot)$. ■

The identification result generalizes ideas in Robinson (1988), who deals with partially linear models where $f(\cdot;\beta)$ is linear. Condition (7) is a conventional identification assumption used in the context of (nonlinear) least squares applied to the parametric function $m(\cdot;\beta)$, which can be obtained from $f(\cdot;\beta)$ and the observable distribution of x_{it} given z_{it}. It would fail, for instance, if expectations (in z_{it}) are static and injective functions *solely* of inputs (in x_{it}), but likely to hold in more general settings where expectations depend on other elements in the information set Ω_{it}. In fact, if $f(\cdot;\beta)$ is linear in parameters (e.g., Cobb-Douglas and translog), the result boils down to that in Robinson

(1988):

Corollary 1. *Under the assumptions of Theorem 1, if $f(\cdot;\beta)$ is linear in parameters and instead of Condition (7) suppose that*

$$\mathbb{E}\left\{[x_{it} - \mathbb{E}(x_{it}|z_{it})][x_{it} - \mathbb{E}(x_{it}|z_{it})]^\top\right\} \tag{8}$$

is non-singular. Then the parameter vector $\theta_0 = (\beta_0, g_0)$ is identified.

Proof. Since $f(\cdot;\beta)$ is linear in parameters we can represent it as $f(x_{it};\beta) = x_{it}^\top\beta$. The result obtains as Condition (8) implies Condition (7). Suppose that there exists $\beta \neq \beta_0$ such that

$$\begin{aligned}&\mathbb{P}\left(f(x_{it};\beta_0) - \mathbb{E}\big[f(x_{it};\beta_0)|z_{it}\big] = f(x_{it};\beta) - \mathbb{E}\big[f(x_{it};\beta)|z_{it}\big]\right)\\ &= \mathbb{P}\left([x_{it} - \mathbb{E}(x_{it}|z_{it})]^\top(\beta_0 - \beta) = 0\right) = 1.\end{aligned}$$

This then implies that

$$\mathbb{E}\left\{[x_{it} - \mathbb{E}(x_{it}|z_{it})][x_{it} - \mathbb{E}(x_{it}|z_{it})]^\top\right\}(\beta_0 - \beta) = 0.$$

This means that $\beta_0 - \beta \neq 0$ is in the nullspace of $\mathbb{E}\left\{[x_{it} - \mathbb{E}(x_{it}|z_{it})][x_{it} - \mathbb{E}(x_{it}|z_{it})]^\top\right\}$ thus implying that this matrix is singular. Hence, Condition (8) implies Condition (7) and the result follows from Theorem 1. ■

This can also be obtained by directly applying the results in Robinson (1988). As discussed there (see p.940), Condition (8) prevents any element of x_{it} (inputs, in our case) from being almost surely perfectly predictable by z_{it} (expectations, in our case) in the least squares sense, although it does not preclude (nonlinear) functional relations among x_{it} elements and identification is possible even if x_{it} uniquely defines z_{it}, when the converse is not true. If x_{it} (inputs, in our case) and z_{it} (expectations, in our case) are observable, perfect predictability of the former by the latter is in principle testable. This condition would fail, as already mentioned above, if expectations about future output and inputs (which appear in z_{it}) are static and injective functions *solely* of inputs (which appear in x_{it}), but likely to hold in more general settings where expectations depend on other elements in the information set Ω_{it} for example.

2.2 Estimation

As highlighted in the previous section, the protocol we put forward builds on the idea that expectations about future output and inputs contain information about future productivity and, under Markovianity, about current productivity. This can be leveraged to proxy for current productivity using observed expectations and thus address econometric endogeneity concerns. Equation (5) can then be used to recover estimates of (θ, g) that are either fully or semi-parametric depending on whether one specifies the functional form of $g(\cdot)$. In the remainder of this paper, we follow a semi-parametric approach, which allows us to avoid imposing structure on $g(\cdot)$ and yields a novel estimation methodology. This section focuses on the case of Cobb-Douglas production technology to outline our proposed methodology, although it generalizes to other specifications (e.g. translog), which we examine in our empirical application.

Cobb-Douglas production implies:

$$\begin{aligned} y_{it} &= \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \epsilon_{it} \\ &= \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \Psi\left(\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \beta_0 - \beta_k k_{it+1} - \beta_l \mathbb{E}[l_{it+1}|\Omega_{it}]\right) + \epsilon_{it}. \end{aligned} \tag{9}$$

Assuming Ψ is a smooth function, equation (9) is an example of a generalized additive model, early explorations of which were provided for instance by Hastie and Tibshirani (1986), and the partially linear model studied by Robinson (1988) among others. Hastie and Tibshirani characterize the non-linear part of the model – in our case, the Ψ function – as a weighted sum of unspecified smooth functions, the parameters and weighting of which can be recovered using (quasi-)maximum-likelihood-based estimation.[9]

There are two specific features of the model of equation (9) that depart from the standard generalized additive model. First, we know that Ψ is monotonic, which amounts to imposing constraints on derivatives of the smooth functions that comprise Ψ. The exact form of these constraints and the consequent optimization problem are derived by Pya and Wood (2015), who also present an algorithm to estimate such 'shape constrained' general additive models that we deploy in our empirical application.

Second, the argument of the smooth Ψ function is itself a function of the model's parameters. To address this issue, we take inspiration from Friedman and Stuetzle (1981), who develop an iterative 'backfitting' algorithm that recovers parameter estimates in additive models where the arguments of the smooth functions are linear functions of parameters (see also Ichimura and Todd (2007)). Adapting the algorithm to our setting

9. As in a linear regression, maximum likelihood based on normal errors amounts to least squares minimization here.

yields the following iterative estimation procedure:

1. Pick initial parameter values $(\hat{\beta}_{k0}, \hat{\beta}_{l0})$.
2. For iteration j, calculate $Z_{ij} = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \hat{\beta}_{kj-1}k_{it+1} - \hat{\beta}_{lj-1}\mathbb{E}_{it}[l_{it+1}|\Omega_{it}]$.
3. Fit the model $y_{it} = \beta_k k_{it} + \beta_l l_{it} + \Psi(Z_{ij}) + \epsilon_{it}$ using the shape constrained estimation protocol of Pya and Wood (2015) to obtain $(\hat{\beta}_j, \hat{\Psi}_j)$.[10]
4. Calculate the Euclidean distance between $(\hat{\beta}_{kj}, \hat{\beta}_{lj})$ and $(\hat{\beta}_{kj-1}, \hat{\beta}_{lj-1})$. If the distance is below some tolerance level, stop and treat $\hat{\beta}_j$ as the model's parameter estimates. If not then update the iteration number $j \leftarrow j+1$ and repeat from step 2.

For more general production functions, one instead should use $Z_{ij} = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1}; \hat{\beta}) dF_{it}(l_{it+1}))$ in step 2 and $y_{it} = f(k_{it}, l_{it}; \beta) + \Psi(Z_{ij}) + \epsilon_{it}$ in step 3. General treatments for the convergence of related procedures is examined, for example, in Pastorello et al. (2003) and Dominitz and Sherman (2005) and we explore theoretical conditions for convergence in Appendix B. We henceforth refer to this iterative algorithm as 'NPR' and provide further implementation details in Appendix C.

2.3 Comparison with other methods

Given the range of existing production function estimation methods, we emphasize three aspects that distinguish our NPR estimation algorithm.

First, unlike the widely-used proxy variable approaches of Olley and Pakes (1996), Levinsohn and Petrin (2003) and Ackerberg et al. (2015) (henceforth OP, LP and ACF respectively), it does not require that firm input decisions be invertible.[11] To see this note that the Ψ function in equation (5) plays a role analogous to the function Φ used by OP to represent firms' investment policy, or LP and ACF to represent firms' material input choice. In OP, LP and ACF, Φ is a function of current inputs used to control for current

10. The constant $\hat{\beta}_{0j}$ is not included in the calculation of Z_{ij} as any constant term in the smooth function cannot be separately identified from the constant term of the linear part of the model.

11. In standard models, assumptions over the information set and optimality of input choices generate the key econometric assumptions that the input demand equation is strictly monotonic in productivity and thus invertible when there is one scalar persistent unobservable. That said, even if the decision rules are not the solution to an optimization problem, as long as invertibility holds one can generate a proxy for ω. This invertibility fails if not all determinants of the firm's decision are observed by the econometrician. We see it as an advantage of our approach that we impose assumptions on a primitive object rather than an equilibrium outcome.

ω. The success of this approach therefore hinges on the existence of a monotonic relationship between contemporaneous productivity and inputs, which is typically assumed with recourse to models of optimal firm behaviour that imply such relationships are monotonic (though it may still hold even if decision rules are not the solution to an optimization problem). NPR, by contrast, requires the assumption that firms' expectations align with the true production technology but allows one to remain agnostic about how firms make their input decisions.[12] As confirmed by the Monte Carlo simulations discussed in section 3, this distinction means NPR remains consistent when firms' decisions are subject to optimization error or when there are additional unobservable variables influencing input decision, whereas other proxy variable methods do not.[13] Some papers have extended proxy methods to include additional variables (e.g., De Loecker and Warzynski 2012 and Eslava et al. 2026).

A second point of distinction is that NPR can accommodate non-linear productivity dynamics, whereas the 'dynamic panel' methods of Blundell and Bond (1998) and Blundell and Bond (2000) typically require linearity. Such non-linearity is enabled by the flexible form of the Ψ function at the core of the NPR method, although it is worth noting the monotonicity constraint required by NPR demands that ω follow a first order Markov process. In theory, a relative strength of dynamic panel methods is that they can be used in situations where ω follows a Markov process of higher order, but in practice this requires the researcher to correctly specify both the AR and MA components of the linear productivity process and requires a longer, and hence more selected, data panel. In practice, the use of longer lags as instruments typically generates estimation problems, especially on the capital coefficient as the series is highly persistent.

The final point is that NPR can identify the production function parameters β and the transition law (g) from a single cross section of data. While repeated observations of current-period and next-period inputs and outputs would accommodate more general models than that presented in section 2.1, such as a production function with firm fixed effects,[14] for this baseline – which is standard in the literature – a single observation per firm is adequate. Both the proxy variable and dynamic panel approaches, by contrast,

12. While the NPR algorithm requires firms' expectations align with the true production technology, the moment condition of equation (5) may still yield a consistent estimator for β in contexts where this does not hold owing to bias in firms' expectations. This is discussed in section 2.5.

13. This feature also favours NPR over 'index number' methods discussed by Van Biesebroeck (2007), such as those proposed by Solow (1957) and Hall (1988), which derive equations expressing production function parameters as functions of observables under the assumption of optimal firm behaviour.

14. See, for instance, Arellano et al. (2024) for an elaboration on this point in the context of earnings dynamics where expectations data helps resolve important issues in dynamic panel data models, such as "Nickell bias".

require multiple observations per firm. Methods do exist for correcting for the selection problem that such sample restrictions introduce, but we regard the absence of any such requirement for NPR as highly attractive.[15]

A disadvantage of the NPR approach is that it requires data on firms' subjective expectations. Since the vast majority of production functions are estimated in logs, we require information of firms' subjective expectation *distributions* because a single value of firms' expected output, for example, would be insufficient to recover firms' expected *log* output. However, questions that provide such information are increasingly being included in firm surveys such as the UK Management and Expectations Survey (MES) (Office For National Statistics 2022) that we use in our empirical application below. We offer further considerations on the emerging data landscape in Section 4. As such data become increasingly available, we believe the three features of NPR discussed above bring notable advantages that warrant its addition to the established suite of production function estimators.[16]

2.4 Accommodating measurement error in expectations data

The expectations data that enable the proposed NPR estimator are, as with all data, susceptible to measurement error. Similarly to the case of measurement error in materials discussed in Ackerberg et al. (2015), measurement error in firms' subjective expectations will violate the invertibility condition required by NPR and thereby undermine the estimator's consistency. While this is true of the baseline NPR estimator outlined above, we can draw on the insights of Evdokimov and Zeleneev (2025) (henceforth EZ) to develop an alternative estimator, "EZ-NPR", which we show below is robust to moderate levels of measurement error in expected output and employment.

The key insight of EZ is that non-linear moments can be "corrected" for measurement error by taking Taylor expansions of the moments with respect to the potentially mismeasured variables around the point of zero measurement error. To demonstrate the application of this in our context, consider the moment conditions that follow from

15. Blundell and Bond (1998) discuss how selection may be controlled for by a firm fixed effects and Olley and Pakes (1996) focus on a proxy variable approach. But the absence of an external instrument in the selection equation may pose identification issues for these approaches.

16. The MES and several other surveys also collect firms' expectations about aggregate, macroeconomic variables (e.g. Survey on Inflation and Growth Expectations by the Bank of Italy or the Business Tendency Survey Dovern et al. 2023). It is possible that those data offer additional, complementary information that can possibly be used as well for the estimation of production functions using moments that aggregate across firms.

equation (5) in the case of Cobb-Douglas production[17]

$$M_{it}^{x}(Z_{it}) = \left(y_{it} - \beta_k k_{it} - \beta_l l_{it} - \Psi\left(Z_{it}\right)\right) x_{it}, \tag{10}$$

where x_{it} is any variable in firm-i's information set Ω_{it} and

$$Z_{it} = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \beta_k k_{it+1} - \beta_l \mathbb{E}_{it}[l_{it+1}|\Omega_{it}].$$

Under the assumptions outlined in Section 2.1, such moments are zero in expectation and can be used to recover unbiased estimates of β. If expectations data are subject to measurement error, however, then

$$\begin{aligned} Z_{it} &= \mathbb{E}^{*}_{it}[y_{it+1}|\Omega_{it}] + \upsilon_{yit} - \beta_k k_{it+1} - \beta_l(\mathbb{E}^{*}_{it}[l_{it+1}|\Omega_{it}] + \upsilon_{lit}) \\ &= Z^{*}_{it} + \upsilon_{yit} - \beta_l \upsilon_{lit}, \end{aligned}$$

where asterisks denote true values and υ_y and υ_l represent measurement error in expected output and labour, respectively. In general, this means that equation (10) will not equal zero in expectation when evaluated at the true parameters using observed data, and we lose identification.

To progress, we follow EZ and take a second-order Taylor expansion of $M_{it}^{x}(Z_{it}) = M_{it}^{x}(Z^{*}_{it} + \upsilon_{yit} - \beta_l \upsilon_{lit})$ around the point $(\upsilon_{yit}, \upsilon_{lit}) = 0$

$$\begin{aligned} M_{it}^{x}(Z_{it}) &= M_{it}^{x}(Z^{*}_{it}) + \frac{\partial M_{it}^{x}(Z^{*}_{it})}{\partial \mathbb{E}_{it}[y_{it+1}]}\mathbb{E}[\upsilon_y] + \frac{\partial^2 M_{it}^{x}(Z^{*}_{it})}{\partial \mathbb{E}_{it}[y_{it+1}]^2}\frac{\mathbb{E}[\upsilon_y^2]}{2} \\ &\quad + \frac{\partial M_{it}^{x}(Z^{*}_{it})}{\partial \mathbb{E}_{it}[l_{it+1}]}\mathbb{E}[\upsilon_l] + \frac{\partial^2 M_{it}^{x}(Z^{*}_{it})}{\partial \mathbb{E}_{it}[l_{it+1}]^2}\frac{\mathbb{E}[\upsilon_l^2]}{2} + \frac{\partial^2 M_{it}^{x}(Z^{*}_{it})}{\partial \mathbb{E}_{it}[y_{it+1}]\partial \mathbb{E}_{it}[l_{it+1}]}\mathbb{E}[\upsilon_y \upsilon_l] \\ &= M_{it}^{x}(Z^{*}_{it}) + \frac{\partial^2 M_{it}^{x}(Z^{*}_{it})}{\partial \mathbb{E}_{it}[y_{it+1}]^2}\frac{\mathbb{E}[\upsilon_y^2]}{2} + \frac{\partial^2 M_{it}^{x}(Z^{*}_{it})}{\partial \mathbb{E}_{it}[l_{it+1}]^2}\frac{\mathbb{E}[\upsilon_l^2]}{2}, \end{aligned} \tag{11}$$

where the second equality assumes that both measurement errors are mean-zero and uncorrelated with one another. EZ show that, up to higher-order terms, the derivatives of the moments evaluated at observed values are consistent estimators of the derivatives evaluated at true values, and thus corrected moments can be constructed as

$$M_{it}^{x}(Z^{*}_{it}) = M_{it}^{x}(Z_{it}) - \frac{\partial^2 M_{it}^{x}(Z_{it})}{\partial \mathbb{E}_{it}[y_{it+1}]^2}\frac{\mathbb{E}[\upsilon_y^2]}{2} - \frac{\partial^2 M_{it}^{x}(Z_{it})}{\partial \mathbb{E}_{it}[l_{it+1}]^2}\frac{\mathbb{E}[\upsilon_l^2]}{2}. \tag{12}$$

17. We focus on the Cobb-Douglas example for ease of exposition, although note that the general EZ approach can be extended to other functional forms.

Treating the variance terms $\mathbb{E}[\upsilon_y^2]$ and $\mathbb{E}[\upsilon_l^2]$ as additional parameters to estimate, one can therefore use corrected moments of the form (12) to recover estimates that are robust to moderate levels of measurement error in firms' expectations.

To understand the meaning of "moderate" in the preceding sentence, observe that the expectations data enter the moment in equation (10) via the uni-dimensional Z_{it} variable, which represents $\mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}]$. The extent to which measurement errors in $\mathbb{E}_{it}[y_{it+1}|\Omega_{it}]$ and $\mathbb{E}_{it}[l_{it+1}|\Omega_{it}]$ both impede performance of the NPR estimator and can be withstood by EZ-NPR therefore depends on how their combined variance (equal to $\mathbb{E}[\upsilon_y^2]+\beta_l^2\mathbb{E}[\upsilon_l^2]$) compares to the variance of the true $\mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}]$. If persistent productivity follows an AR(1) process, for example, NPR and EZ-NPR can tolerate measurement error in firms' expectations of increasingly higher variance as the persistence of the process increases, since this implies greater variance in $\mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}]$. We explore this quantitatively in Monte Carlo simulations discussed in Section 3.

At first blush, EZ-NPR appears simpler to implement than the iterative NPR algorithm: consistent estimates can be obtained via GMM estimation using corrected moments of the form (12). In practice, however, the non-linear high-dimensional objective function proved hard to optimise numerically. We therefore concentrate out the spline and variance parameters to focus optimisation on the two-dimensional production function parameter vector of interest. This necessitates nesting an iterative procedure within the EZ-NPR objective function, which we discuss alongside further implementation details in Appendix D.2.

2.5 Accommodating biased expectations and imperfect knowledge of the production technology

In our baseline case, when firms know the true production technology and their beliefs align with this, the law of motion for ω_{it} can be inverted to obtain equation (4). This inversion is crucial to the NPR estimation algorithm and is analogous to the invertibility condition that OP impose on the investment policy function and that LP and ACF impose on the material input policy function. We thus rely on a one-to-one mapping between firms' expectations and their current productivity to generate a proxy for the latter.

Given the centrality of expectations to the NPR algorithm, it is important to consider whether and how expectational biases undermine the proposed approach. The first thing to note is that our suggested method can accommodate biased input expectations as long as such bias is also reflected in firms' expected output and vice-versa. If, for example, a

firm is systematically optimistic in its sales forecasts we would require it to be similarly optimistic in its employment forecasts. The precise meaning of 'similarity' in this context is governed by the production function. Specifically, in the case of a firm with over-optimistic output expectations, we require bias in the firm's employment expectations such that the integral of the production function with respect to expected labor equals the biased output expectation.

As an example, consider the case of Cobb-Douglas production in equation (9). When the bias in sales expectations, say $\texttt{bias}_{y,it}$, balances the bias in the employment expectations, say $\texttt{bias}_{l,it}$, so that $\texttt{bias}_{y,it} = \beta_l \texttt{bias}_{l,it}$, then one is still able to recover an unbiased expectation of firms' next-period productivity as the residual on the right hand side of equation (4). Furthermore, in this particular case, when production is Cobb-Douglas, it is possible to construct a 'Wald'-type estimator of β_l if one has the data to compare expectations of output and labor to their realized values. We do not pursue this in our main analysis because we present evidence that such bias is very minimal in our empirical context (see the discussion of Table 7 in subsection 4.2 and also Bloom et al. 2026). Nevertheless, we provide a further exposition in Appendix D.

Since the vast majority of literature on managerial bias has focused on biases in output expectations, the ability of our method to accommodate this is encouraging. Biases in firms' productivity expectations, however, are more problematic. To see why, suppose firms' bias about their next-period productivity is captured by ι_{it} (positive values reflect optimism, negative values reflect pessimism), such that

$$\mathbb{E}[\omega_{it+1}|\Omega_{it}] = g(\omega_{it}) + \iota_{it}.$$

Even if firms' expectations about output and inputs align with the true production technology, the presence of bias means

$$\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] = \int f(k_{it+1}, l_{it+1}; \theta) dF_{it}(l_{it+1}) + g(\omega_{it}) + \iota_{it}. \tag{13}$$

The bias term ι_{it} therefore violates the strict monotonicity assumption we require to recover ω since

$$g(\omega_{it}) + \iota_{it} = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1}; \theta) dF_{it}(l_{it+1})$$

$$\therefore$$

$$g^{-1}\left(\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1}; \theta) dF_{it}(l_{it+1})\right) = g^{-1}\left(g(\omega_{it}) + \iota_{it}\right) \neq \omega_{it}.$$

Whether such bias is surmountable depends on its form. The simplest case, which we are able to test in our empirical application, is when the bias term ι_{it} is uncorrelated with all observables used by the NPR estimator (i.e. $\{y_{it}, l_{it}, k_{it}, E_{it}[y_{it+1}], E_{it}[l_{it+1}], k_{it+1}\}$). In this case, the bias term operates exactly as the measurement error discussed in Section 2.4 and consistent estimates can be recovered using the EZ-NPR estimator, which constructs instruments that are "corrected" by bias via a second-order Taylor expansion (Evdokimov and Zeleneev 2025).

In addition to this type of "classical" bias, we discuss in Appendix D how the NPR estimator can be extended to withstand bias that is either time-invariant or a function of observables. In these cases, one can embed the NPR estimation algorithm in an outer iterative estimation loop to recover estimates of firms' expected productivity bias, which can subsequently be used to restore consistency of the NPR estimator. These methods require panel data on expectations which are very limited data in our empirical application, so we do not pursue them further here.[18]

Another factor that undermines consistency of the NPR estimator is if the firm has imperfect knowledge of the production technology. In this case, deviations between the production function parameters perceived by firms and the true values create a 'wedge' between expected output, the true production function evaluated at $(k_{it+1}, \mathbb{E}_{it}[l_{it+1}])$ and $g(\omega_{it})$, similar to the bias term ι_{it} of equation (13). In Appendix D we demonstrate an extension to the baseline NPR estimator that recovers consistency in the presence of such imperfect knowledge. However, similarly to the bias-robust extensions this requires a panel of firms' expectations.

3 Monte Carlo simulations

Our baseline Monte Carlo setup follows that of ACF (see more details in Appendix E). The production function specification is Leontief in the material input:

$$Y_{it} = \min\{\beta_0 K_{it}^{\beta_k} L_{it}^{\beta_l} e^{\omega_{it}}, \beta_m M_{it}\} e^{\epsilon_{it}},$$

where $\beta_0 = 1, \beta_K = 0.4, \beta_l = 0.6$ and $\beta_m = 1$. In our baseline analysis the productivity shock is assumed to follow an AR(1) process:

$$\omega_{it} = \rho \omega_{it-1} + \xi_{it}, \tag{14}$$

18. Although we have a reasonable panel for outputs and inputs, the MES subjective expectations data are two cross sections with limited longitudinal overlap.

with $\rho = 0.7$. As pointed out by ACF, the LP estimator does not identify the production function parameters unless there is stochastic variation in firms' labor inputs, for example due to optimization error. We therefore focus on data generating processes featuring such variation, which we introduce in the same way as ACF by adding a mean-zero normally-distributed random variable to firms' optimal level of labor. In addition, we also consider the impact of optimization error in investment and materials (which ACF do not consider)[19], which we incorporate in the same manner as labor.[20]

For each DGP, we use the closed-form solutions of the model to simulate data for 1,000 firms over 100 periods. Capital is initialised at zero and we only use data from the last 10 periods for estimation purposes, as by this time the capital stock appeared to have reached steady state. Further details of the environment and the data generating process (DGP) are given in Appendix E (and the Appendix in Ackerberg et al. (2015)). The ACF estimator involves minimization of a nonlinear objective function over a two-dimensional parameter space. We use the R `optimx` function to find its minimum, and this requires the researcher to specify initial values of the parameters. Similarly to the NPR algorithm, we implement the ACF estimator using 16 initialization points (the interaction of two equally-spaced 4-point grids on a 0.1 to 0.9 range - one for the initial labor coefficient value and one for the initial capital coefficient value), and retain the estimates that return the lowest value of the objective function. The nonlinear optimization component of the OP and LP estimators, by contrast, is unidimensional and the method we use to find the minimum (the R function optimize which, unlike optimx, is a unidimensional optimizer) instead requires the researcher to specify a search grid, which we set to [0,2]. Experimentation indicates that estimators are generally robust to varying these optimization settings. The only exceptions arose when using initialization grids solely consisting of points close to 0 or 1, in which case the ACF estimator failed to return any estimates within the 0-1 range.

Table 1 examines the performance of the estimators as various firm choices are subject to optimization error. We highlight three salient points. First, as expected, the OLS and OP parameter estimates are heavily biased across all DGPs. Bias in the OP

19. ACF's analysis considers the impact of *measurement* error in materials but this is distinct from optimization error as it does not affect output. Optimization error, by contrast, will affect output via the assumption of Leontief technology.

20. Optimization errors in labor and investment are simulated from a mean-zero normal distribution with standard deviation 0.37, which matches the distributional assumption ACF make regarding the labor optimization error in their DGPs. By contrast, we simulate the optimization error in materials from a mean-zero normal distribution with a standard deviation of 0.185 (i.e. half of 0.37). We do this because the ACF estimator is vulnerable to this type of optimization error and failed to return any plausible estimates when simulations used the higher standard deviation.

estimates is due to the presence of firm-specific capital adjustment costs, added by ACF to obtain across-firm variation in capital similar to that observed in the data. Second, when optimization error affects labor only (first Panel A), LP and NPR all perform well, as does ACF when disregarding implausible estimates (i.e. when the values of the output elasticities are below zero or greater than one).[21] As anticipated by ACF, LP (slightly) outperforms their proposed estimator in this environment in terms of precision. NPR improves on LP even more, achieving much greater precision on the capital coefficient.

21. Those implausible estimates relate to a global identification issue in the moments employed by ACF:

> "...there is a identification caveat using our suggested moments in all three of these DGPs. More specifically, there is a "global" identification issue in that the moments have expectation zero not only at the true parameters, but also at one other point on the boundary of the parameter space where $\beta_k = 0$ and $\beta_l = \beta_l + \beta_k$, and the estimated AR(1) coefficient on ω equals the AR(1) coefficient on the wage process. One can easily calculate that at these alternative parameter values, the second stage moment equals the innovation in the wage process, which is orthogonal to k_{it} and l_{it-1}. This "spurious" minimum is a result of labor satisfying a static first order condition, and we suspect it would not occur were labor to have dynamic implications, nor when the alternative moments (29) are assumed."
>
> Ackerberg et al. (2015), p. 2438, footnote 16.

Table 1: Input Optimization Error Monte Carlo Results

	$\beta_l = 0.6$				$\beta_k = 0.4$				
	Mean	Median	S.D.	MSE	Mean	Median	S.D.	MSE	N runs
	A. Optimization error in l								
NPR	0.600	0.600	0.003	0.000	0.400	0.400	0.005	0.000	500
OLS Levels	0.919	0.919	0.002	0.102	0.098	0.098	0.004	0.091	500
OP	0.840	0.840	0.004	0.057	0.162	0.161	0.008	0.057	500
LP	0.600	0.600	0.003	0.000	0.401	0.400	0.013	0.000	500
ACF	0.823	0.601	1.043	1.135	0.161	0.401	1.125	1.320	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.600	0.600	0.009	0.000	0.401	0.401	0.016	0.000	478
	B. Optimization error in (l, i)								
NPR	0.600	0.600	0.003	0.000	0.400	0.400	0.005	0.000	500
OLS Levels	0.919	0.919	0.002	0.102	0.097	0.097	0.004	0.092	500
OP	0.897	0.897	0.002	0.088	0.105	0.106	0.009	0.087	500
LP	0.616	0.616	0.003	0.000	0.376	0.376	0.012	0.001	500
ACF	0.747	0.606	0.848	0.740	0.237	0.391	0.922	0.875	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.605	0.605	0.010	0.000	0.391	0.392	0.017	0.000	486
	C. Optimization error in (l, m)								
NPR	0.600	0.600	0.005	0.000	0.398	0.400	0.051	0.003	500
OLS Levels	0.919	0.919	0.002	0.102	0.098	0.098	0.004	0.091	500
OP	0.840	0.840	0.004	0.057	0.161	0.161	0.009	0.057	500
LP	0.304	0.304	0.006	0.088	0.770	0.770	0.021	0.137	500
ACF	0.363	0.354	0.186	0.091	0.690	0.697	0.206	0.126	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.355	0.353	0.025	0.061	0.699	0.697	0.032	0.090	499
	D. Optimization error in (l, i, m)								
NPR	0.600	0.600	0.006	0.000	0.399	0.400	0.034	0.001	500
OLS Levels	0.919	0.919	0.002	0.102	0.097	0.097	0.004	0.092	500
OP	0.897	0.897	0.003	0.088	0.104	0.105	0.011	0.088	500
LP	0.304	0.304	0.006	0.088	0.763	0.763	0.020	0.132	500
ACF	0.383	0.347	0.401	0.207	0.664	0.701	0.441	0.264	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.347	0.347	0.028	0.065	0.704	0.701	0.035	0.093	496

Note: number of replications given in the 'N runs' column. The true values of β_l and β_k are 0.6 and 0.4 respectively. l is ln(labor), m is ln(materials), k is ln(capital) and i is ln(investment). The final row in each panel drops ACF estimates where either β_l or β_k are less than zero or greater than one.

Third, when optimization error affects material inputs, all estimators except NPR deteriorate. The LP estimator is vulnerable both to errors in labor and investment simultaneously (second Panel B of Table 1), and labor and materials simultaneously (third Panel C). While ACF appears robust to errors in investment (if one ignores implausible estimates), it is compromised by errors in materials which violate the monotonicity condition ACF require between material input choices and productivity. This is compounded further when optimization errors to labor, investment and materials occur simultaneously, as shown in the final Panel D. In contrast to the other estimators, the NPR estimates

remain both consistent and accurate throughout, although optimization error in materials reduces the precision of the NPR estimates, particularly for the capital coefficient.[22] This is a clear demonstration of the observation made in subsection 2.3 that NPR is robust to optimization error in inputs, whereas the other proxy variable estimators are not.

The results of Table 1 demonstrate the robustness of NPR to optimisation error in contemporaneous firm decisions. Since NPR does not rely on any assumptions related to such decisions, unlike the other proxy variable estimators, this could be regarded as somewhat unsurprising. More compelling tests of the NPR estimator relate to measurement error in expectations and to bias in expected productivity which, as discussed in Sections 2.4 and 2.5, violate consistency of the NPR estimator.

Table 2 shows moments of parameter estimates obtained by applying the NPR estimator to data simulated under the 'optimization error in l' scenario (Panel A of Table 1) but with the addition of idiosyncratic shocks to firms' expectations. The parameters obtained by all estimators other than NPR are omitted from the table because they do not use the information contained in firms' expectations and are hence almost identical to the old first Panel A. As explained in section 2.5, NPR is robust to biased expectations over labor (the first row). This is because bias in expected inputs leads to bias in expected output according to the production technology, which means that a one-to-one mapping between expected outputs, inputs and productivity is preserved. The second and third rows, however, show that NPR loses consistency when there is bias in expected output or expected productivity. In these DGPs, the relationship between expected outputs and inputs is subject to two unobservables – expected productivity and the bias shock – and hence it is no longer possible to control for expected productivity using expectations data. The fourth row considers a similar DGP to that of the third row, although in this case the bias to firms' expected productivity is a function of a time-variant observable characteristic, $X_{it} \equiv \text{mgmt}_{it}$.[23] Although this type of expectation bias again compromises the performance of the basic NPR estimator, application of a modified version of the NPR estimator described in detail in Appendix D recovers consistency.[24]

22. Despite not relying on material input data, NPR is affected by materials optimization error because of the assumption of Leontief production. When material optimization error is negative, the Leontief assumption means output will be determined by sub-optimally low materials and hence the specification of output as a function of labor and capital will be incorrect.

23. We use the label $mgmt_{it}$ because Bloom et al. (2026) find evidence that forecast biases are related to managerial quality.

24. The bias-robust extension is discussed in 'Case 3' of Appendix D. Management is simulated as a standard normal random variable drawn for each firm-period (i.e. $mgmt_{it} \sim N(0,1)$) and firms' expected productivity bias is simulated as -0.15 times management (i.e. $\iota_{it} = -0.15 mgmt_{it}$).

In summary, NPR is robust to bias in expected inputs and, while bias in expected output or productivity may undermine performance of the basic estimator, extensions to NPR can accommodate these biases under certain assumptions (see Appendix 2.5 for more details). We do not consider the bias-robust versions of NPR further since in our MES data, output expectations appear largely unbiased and we do not have many firms with multiple longitudinal data on firms' subjective expectations.

Table 2: Expectation Bias Monte Carlo Results

		$\beta_l = 0.6$				$\beta_k = 0.4$				
DGP	Estimator	Mean	Median	S.D.	MSE	Mean	Median	S.D.	MSE	N runs
Biased $\mathbb{E}_{it}[l_{it+1}]$	NPR	0.600	0.600	0.003	0.000	0.413	0.408	0.014	0.000	500
Biased $\mathbb{E}_{it}[y_{it+1}]$	NPR	0.918	0.918	0.007	0.101	0.100	0.099	0.006	0.090	500
Biased $\mathbb{E}_{it}[\omega_{it+1}]$	NPR	0.900	0.900	0.002	0.090	0.124	0.124	0.006	0.076	500
Biased $\mathbb{E}_{it}[\omega_{it+1}]$ f(mgmt)	NPR	0.684	0.684	0.003	0.007	0.329	0.327	0.050	0.008	500
Bias in $\mathbb{E}_{it}[\omega_{it+1}]$ f(mgmt)	bias-robust NPR	0.600	0.600	0.003	0.000	0.402	0.402	0.007	0.000	500

Note: number of replications given in the 'N runs' column. The true values of β_l and β_k are 0.6 and 0.4 respectively. l is ln(labor), y is ln(output) and ω is persistent productivity. All DGPs feature optimization error in labor.

While the OP/LP and ACF control function estimators do not use subjective expectations data directly, they will be affected by expectational biases if such biases lead to changes in firm decisions. To illustrate this point, we focus on biased productivity expectations and run several simulations changing the variance of the bias component. In contrast to our baseline DGP, we eliminate heterogeneity in investment costs and set the degree of optimization error in labor to 0.1 (instead of 0.37). These changes are made to reduce the bias in the OP estimator in a setting without expectational biases and thus highlight a distinct source of bias.[25]

Figure 1 illustrates how mean squared errors of the control function estimates vary with the magnitude of bias in expected productivity. For small levels of expected productivity bias the NPR estimator returns unbiased results (i.e. a small mean squared error), but its performance deteriorates as the magnitude of the bias grows. This is expected following section 2.5, which explains that expected productivity bias violates the scalar unobservable assumption the NPR estimator depends on. The performance of the other control function estimators is more surprising. The OP estimator performs similarly to the NPR estimator, obtaining unbiased results when expected productivity is subject to low levels of bias but becoming increasingly biased as the magnitude of expected productivity bias grows. The LP and ACF estimators, by contrast, remain unbiased at all levels

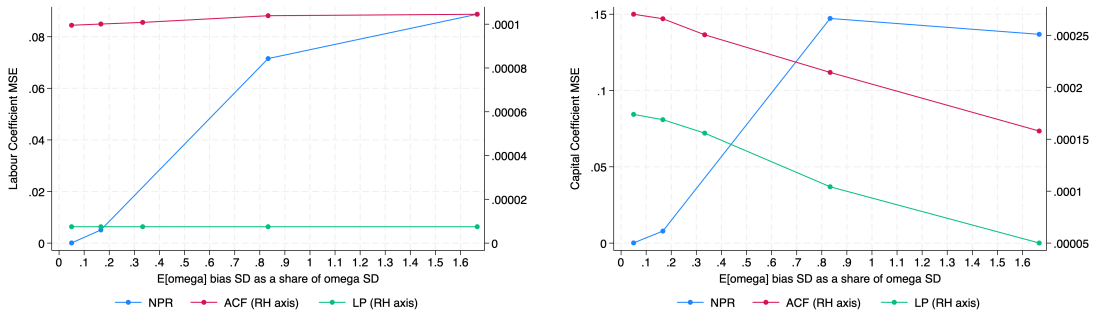
25. Results available from the authors on request show the performance of the LP, ACF and NPR estimators is very similar when productivity bias is varied in the context of the baseline DGP.

of expected productivity bias, with the capital coefficient estimate of both estimators showing small gains in precision as the magnitude of the bias increases.

Figure 1: Estimator MSE Under Different Levels of Expected Productivity Bias

(a) Labor Coefficient MSE

(b) Capital Coefficient MSE



Note: the ACF series shows the mean squared error of ACF estimates across simulation runs after discarding runs where the estimator obtained $\hat{\beta}_l$ or $\hat{\beta}_k$ less than zero or greater than 1.

The relative performance of the OP/LP/ACF estimators has a straightforward explanation. Firm-specific biases in expected next-period productivity will increase between-firm differences in current-period investment, leading to increased variation in next-period capital. Conditional on capital (and, in the case of ACF, labor) materials is still chosen optimally every period which, in our DGPs in Table 2, generates a one-to-one mapping between materials (conditional on other inputs) and productivity. This preserves the monotonicity assumption that is crucial to both the LP and ACF estimators. Biased expectations of future productivity therefore leave the consistency of the LP/ACF estimators intact, while also increasing variation in capital, which translates into greater precision in these estimators. In contrast, the extra variation that biased productivity expectations introduce into investment violates the scalar unobservable assumption that the OP estimator relies on. The same logic applies when labor or output expectations are subject to bias, leading to similar results as those shown in Figure 1. Thus, an additional benefit of the LP and ACF estimators over OP is that, by using conditional demand for a static input to control for persistent productivity, they are robust to firm-specific variation in expectation accuracy. Such firm-specific variation will likely influence dynamic decisions such as investment, violating the scalar unobservable assumption required for unbaisedness of the OP estimator.

4 Data

4.1 Sources and sample characteristics

Our proposed methodology requires data on firms' log output, log inputs and one-period-ahead expectations of these quantities. While numerous firm surveys include questions related to expectations (e.g., Table 2.1 in Born et al. (2022)), the majority of these ask firms for single point-valued expectations of future levels or growth rates. On its own, this is not enough to obtain measures of firms' expectations of *log* levels. To do so, one requires further information on firms' subjective expectation distributions, such as variance or, ideally, several points and associated probabilities from firms' subjective cumulative distribution functions (CDFs).

Table 3 documents surveys that contain enough information to estimate firms' expected log quantities and thus implement our proposed estimator or that would require small modifications to do so. These include the Management and Organizational Practices Surveys (MOPS) (Buffington et al. 2016) by the US Census Bureau, the Decision Maker Panel (DMP) by the Bank of England (Bloom et al. 2017), the Survey of Business Uncertainty (SBU) by the Atlanta Federal Reserve Bank and the China Employer-Employee Survey (CEES) (Altig et al. 2022).

Several points are worth highlighting on Table 3. First, it is possible to implement our proposed estimator in three of the six largest economies (in the world in 2024), including the US and China. Small modifications to existing surveys would additionally include Japan, Germany and Italy. The estimator outlined above thus has broader relevance than one might initially presume.

Second, the most common method of eliciting firms' expectations is to ask for expected outcomes in a number of quantitative 'scenarios' (e.g. 'low', 'medium' and 'high'), along with perceived likelihoods of these scenarios occurring. Section 4.2 discusses how to recover measures of firms' subjective expectation distributions from such information, which researchers using these datasets may find useful beyond the subject of production function estimation.

A final remark on Table 3 is that the number of surveys eliciting information on firms' subjective expectation *distributions* is smaller than those that elicit information on point-valued expectations. Born et al. (2022), for example, document a further 13 surveys that contain questions on point-valued expectations. This suggests data collection authorities have strong interest in measuring firms' expectations, yet the manner in which they conventionally do so substantially constrains the analytic utility of such information. Our paper is one example of the econometric possibilities enabled by data on

Table 3: Surveys including questions on firms' subjective expectation distributions

Country	Survey	Documentation	Sectors	Freq.	Years avail.	Years NPR feas.	Turnover[1]	Employment	Expectations	Extra Qs for NPR	Example publication
NPR feasible											
AU	Business Outlook Scenarios Survey (BOSS)	https://tinyurl.com/5fx7manu	Manuf. and services	Monthly	Jan 2025–present	Jan 2025–present[2]	t, $\mathbb{E}[\frac{t+1}{t}]$	t, $\mathbb{E}[t+1]$	5 scenarios with probabilities	NA	https://www.nber.org/papers/w34836
CN	China Employer–Employee Survey	https://tinyurl.com/2p8yp6xz	Manuf.	Annual / Biannual	2015, 2016, 2018	2018	$t-1$, t, $\mathbb{E}[t+1]$	$t-1$, t, $\mathbb{E}[t+1]$	5 scenarios with probabilities	NA	
UK	Management and Expectations Survey	https://tinyurl.com/y5uvawdk	Manuf. and services	Triannual	2017, 2020, 2023	2017, 2020, 2023	$t-1$, t, $\mathbb{E}[t+1]$	$t-1$, t, $\mathbb{E}[t+1]$	5 scenarios with probabilities	NA	
UK	Decision Maker Panel	https://tinyurl.com/y46e4wmk	Manuf. and services	Monthly	2016–present	2016–present[2]	t, $\mathbb{E}[\frac{t+1}{t}]$	t, $\mathbb{E}[\frac{t+1}{t}]$	5 scenarios with probabilities	NA	https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2024/the-decision-maker-panel-a-users-guide.pdf
US	Survey of Business Uncertainty	https://tinyurl.com/mv5vpnh8	Manuf. and services	Monthly	2014–present	2014–present	$t-1$, t, $\mathbb{E}[t+1]$	$t-1$, t, $\mathbb{E}[t+1]$	5 scenarios with probabilities	NA	https://www.sciencedirect.com/science/article/pii/S0304407620302785#sec1
US	US Management and Organizational Practices Survey (MOPS)	https://tinyurl.com/mt2aurz8	Manuf.	Approx. every 5 years	2010, 2015, 2021	2015, 2021	$t-1$, t, $\mathbb{E}[t+1]$	$t-1$, t, $\mathbb{E}[t+1]$	5 scenarios with probabilities	NA	https://www.nber.org/papers/w28259
NPR feasible with slight modifications											
DE	Bundesbank Firm Expectations Survey	https://tinyurl.com/y4dssbp5	Manuf. and services	Monthly	2020–2024		t, $\mathbb{E}[t+1]$ (2021–2025)	t, $\mathbb{E}[t+1]$ (2024Q3 only)	Turnover: 5 scenarios with probabilities; Employment: mode with bounds	Add scenario-based employment expectations	
IT	Survey of Industrial and Service Firms (INVIND)	https://tinyurl.com/5n6z54av	Manuf. and services	Annual	1996–present		$t-1$, t, $\mathbb{E}[t+1]$	$t-1$, t, $\mathbb{E}[t+1]$	Turnover (2017–2020): mean with probability bins; Employment: point forecast	Add subjective employment distribution	https://link.springer.com/article/10.1007/s00181-025-02725-0
JP	Japanese MOPS (JP MOPS)	https://tinyurl.com/nhka5723	Manuf.	Variable	2017, 2021		$t-1$, $\mathbb{E}[t]$[2], $\mathbb{E}[t+1]$	$t-1$, $\mathbb{E}[t]$[3], $\mathbb{E}[t+1]$	$\mathbb{E}[t]$: point; $\mathbb{E}[t+1]$: 3 scenarios with probabilities	Add distribution over current $\mathbb{E}[t]$ or administer survey closer to year end	https://www.sciencedirect.com/science/article/pii/S0889158321000319?via%3Dihub

Notes. (1): "Turnover" refers to sales revenue. (2): questionnaire has a rotating structure, with expectations questions for turnover and employment asked in distinct months. Implementing NPR would therefore require an assumption about the stability of employment across months. (3): the JP MOPS was administered between January and March of the survey year. This is unlike the other surveys, which were typically administered towards the end of the survey year. Current-year values are therefore subject to far greater uncertainty in the JP MOPS than in other datasets and it is therefore inappropriate to use point-valued reports to infer log quantities.

subjective expectation distributions and we hope it motivates further collection of such data. Point-valued expectations can be insightful but far greater insight can be gained from more detailed expectation measures, even if the additional detail is only some quantification of uncertainty/variance. Furthermore, this paper demonstrates how questions on expectations enable analysis that typically demands panel data to be conducted using a cross section; a very attractive feature given the costs of panel data collection. Including questions on subjective expectation distributions in cross-sectional surveys, such as the World Bank Enterprise Survey, would therefore enable production function estimation in contexts where prominent methods are currently infeasible due to an absence of firm-level panel data.

Our empirical application uses data from the Management and Expectations Survey, one of the datasets in Table 3. MES is a survey administered by the UK's statistical authority, ONS, and is sent to a stratified random sample of non-financial private sector firms.[26] The MES was designed to have broadly the same set of questions as the Atlanta Fed's Survey of Business Uncertainty and US Management and Orangizational Practices Survey (US MOPS, see Bloom et al. (2019) for details) and was administered in 2017 and 2020, creating two 'waves' of data. It can be linked to other administrative business datasets and we exploit this property to match MES respondents to the Annual Business Survey (ABS, Office For National Statistics (2023)) enabling us to compare firms' subjective expectations to outturns.[27] Both the 2017 and 2020 versions of the MES ask firms for their turnover (the British term for sales revenue), employment, capital expenditure and expenditure on intermediates (purchases of energy, materials and services) in the year of the survey and the previous year. The survey specifies it should be completed by 'the most senior person responsible for day-to-day operations', as a senior member of staff who is likely to have adequate knowledge of these quantities (this will correspond to the plant-manager or COO in most firms). We use firms' reported investment in these years to build capital stocks via the perpetual inventory method, imputing base period capital from national accounts data on industry-specific capital stocks, apportioning industry totals among firms according to within-industry intermediate input shares.

The latest 2023 wave of the MES omitted questions on previous-year quantities. Although, as indicated in Table 3, it is feasible to implement our proposed estimator using this data, it is not possible to implement the OP/LP/ACF control-function estimators

26. Formally, these are "Reference Units" which are groups of establishments (e.g., manufacturing plants or retail stores). Further details on the MES' sampling design are provided in Bloom et al. (2026).

27. Establishments with 250 or more employees are surveyed each year by the ABS. Businesses below this threshold are surveyed on a multi-year basis with surveyed businesses in any one year chosen as a stratified random sample from the population.

owing to the absence of previous-year quantities. We therefore exclude this data from our empirical application to ease the comparison of results across estimators.

Inclusion in the subsample of MES data we use in the majority of our analysis is conditioned on responding to all MES expectations questions and having adequate observations of sales, capital, employment, intermediates and investment to implement the OP, LP and ACF estimators. To allow for parameter heterogeneity by sector, we confine attention to three industries defined using the UK's Standard Industrial Classification (SIC): 'electronics' consists of firms that manufacture computer electronic and optical products or electrical equipment (SIC groups 26 and 27 respectively); 'retail' consists of firms in the wholesale and retail trade except motor vehicles and motor cycles (SIC groups 46 and 47 respectively); and 'restaurants' consists of firms who conduct food and beverage serving activities (SIC group 56). These industries were selected as they are among the largest in the MES and provide examples of manufacturing as well as service activities. We have also estimated pooling across all industries.

Table 4 summarizes the characteristics of the MES analysis subsample in terms of means, standard deviation and (fuzzy) median. We show these for the whole MES across all industries (first panel), and the three industries we focus on (subsequent panels). Relative to the combined sample, electronics firms are smaller in terms of inputs and output. Firms in the retail sector are relatively similar in size to the overall non-financial private sector economy, whereas restaurants are larger than average in terms of employment but smaller in terms of sales and value added indicating this sector is relatively labor intensive.

Table 4: MES Analysis Sample Characteristics

	Mean	S.D.	p50	N obs.	N firms
	Full MES				
Turnover (£k)	21672	51818	5550	31508	13460
Employment	154	283	55	31504	13460
Capital (£k)	12590	40543	1567	38028	13457
Capex (£k)	1120	4140	95	33611	13457
Intermediates (£k)	14475	38992	2746	31468	13460
Value added (£k)	7007	18324	1550	31437	13460
	Electronics				
Turnover (£k)	13891	23422	6042	992	425
Employment	84	98	48	992	425
Capital (£k)	6199	11923	2088	1185	425
Capex (£k)	364	907	95	1065	425
Intermediates (£k)	9198	20100	3017	991	425
Value added (£k)	5190	10345	2071	990	425
	Retail				
Turnover (£k)	31137	77214	7732	4429	1869
Employment	118	250	41	4430	1869
Capital (£k)	7513	32134	793	5449	1869
Capex (£k)	820	3136	80	4768	1869
Intermediates (£k)	25522	72175	4714	4426	1869
Value added (£k)	6670	18897	1502	4421	1869
	Restaurants				
Turnover (£k)	11825	22157	2990	949	430
Employment	304	508	81	949	430
Capital (£k)	11005	24882	2354	1135	430
Capex (£k)	1173	3355	75	992	430
Intermediates (£k)	8533	19719	1451	950	430
Value added (£k)	4178	9796	590	949	430

Note: 'Mean' and 'S.D.' columns show the mean and standard deviation calculated over the sample indicated in the 'N obs.' column which consists of a number of distinct firms indicated in the 'N firms' column. While the table is calculated using data from two MES waves, 'N obs.' is often more than twice as large as 'N firms' as the statistics are calculated from firms' reports of current and previous year values, yielding at most two observations per firm-wave. The 'p50' column contains the mean value among the 50 observations closest to the median owing to data disclosure requirements. Monetary values are given in current prices.

4.2 Subjective expectations

To elicit firms' expectations for the following year, the MES asks firms to report on five scenarios ranging from 'lowest' to 'highest'. Firms are asked for the value they

expect each variable separately to take under each scenario in the next year and the likelihood of the scenario occurring. In the 2017 MES, for example, the exact wording of the question regarding turnover (i.e. sales revenue) expectations was: "Looking ahead to the 2018 calendar year, what is the approximate pound sterling value of turnover you would anticipate for this business in the following scenarios [Lowest, Low, Medium, High, Highest], and what likelihood do you assign to each scenario?". This wording is very similar to that used in the US MOPS and in the other surveys listed in Table 3. An image showing the expectations question and its position in relation to questions on current and previous year turnover is given in Appendix F.

The 2017 MES used these questions to elicit firms' expectations of turnover, employment, capital expenditure and expenditure on intermediate inputs (energy, goods and services which we label "materials"), whereas the 2020 MES only asked firms about their expectations of turnover and employment in order to limit survey length. In both years, expectations over monetary quantities were asked in nominal terms and we therefore conduct all analysis on a nominal basis and include a time dummy to allow for industry-year specific shocks (like output prices).

The 2017 MES was administered as a paper survey and, although firms were instructed to ensure the likelihoods assigned to the five scenarios summed to 100, some responses did not meet this criteria. In these cases the reported likelihoods were rescaled to sum to 100 if the total likelihood across the five scenarios was between 90 and 110. A small number of responses with a total reported likelihood outside of this window were discarded. This issue does not appear in the 2020 data as this wave of the MES was administered online and required respondents' reported likelihoods to sum to 100 before they could proceed to subsequent questions.[28]

The MES respondents report point-values for each of the five scenarios and their related probabilities. This is in contrast to the survey design implemented in, for example, Dominitz and Manski (1997), which recovers households' subjective CDFs of one-year-ahead income by asking for the perceived likelihood that income next year will fall below a number of thresholds, where the thresholds are determined by first asking households for the minimum and maximum income they expect next year and splitting the interval into 'bins'. Because of this discrepancy, it is not obvious how to use the MES' questions

28. Both the 2017 and 2020 MES contain a limited number of what appear to be data entry errors in turnover and employment. We identify these by calculating ratios of turnover to employment and comparing these with equivalent ratios observed in the ABS. We identify spurious observations as those whose turnover-employment ratios differ across the MES and the ABS by a factor of two *and* with year-to-year growth in their MES turnover-employment ratio in the top 5% of the distribution. These observations are dropped from our analysis.

on subjective expectations to recover firms' subjective CDFs. One approach is to treat the scenario values as points on either a corresponding CDF or survival function on the stated support. In the CDF approach, for example, the cumulative likelihood for the 'Medium' scenario would be taken as the sum of the likelihoods a firm reports against the 'Lowest', 'Low' and 'Medium' scenarios. In the survival function approach, by contrast, the cumulative distribution function value for the 'Medium' scenario would be taken as one minus the survival function, which is the sum of the likelihoods a firm reports against the 'Highest', 'High' and 'Medium' scenarios.

Experimentation with both approaches found the former (CDF) created a lower mean expectation relative to a simple weighted sum across scenarios, while the latter(survival function) created a higher mean expectation compared with a simple weighted sum across scenarios. We therefore estimate lognormal parameters for both approaches by choosing mean and variance parameters to minimize the sum of squared deviations between the fitted distribution and firms' reported scenario values and their corresponding CDF or survival function points.[29] Firms' subjective CDFs are then characterised as $F_{it}(l_{it+1}) = \mathcal{N}(\bar{\mu}_i, \overline{\sigma_i^2})$, where bars denote the averages across the CDF and survival function estimates.[30]

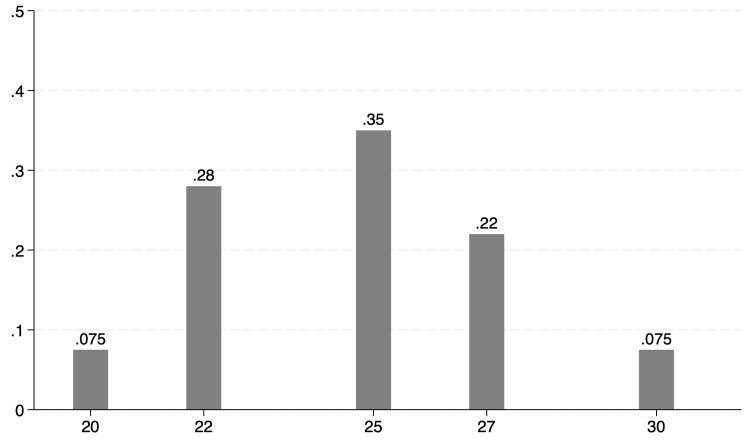
Figure 2 provides an example of our approach. Subfigure 2a is a hypothetical response to a scenario-based expectation question. The values on the x-axis represent the expected values in each of the five scenarios (ranked lowest to highest), and the bars represent the subjective likelihoods of these scenarios occurring. The markers in subfigure 2b show the empirical CDF and survival function implied by these responses, calculated using the method described above. The dashed lines in the subfigure are lognormal CDFs with mean and variance chosen to minimise the sum of square residuals between the fitted CDF and the empirical distribution. The solid line is a lognormal CDF with mean and variance taken as the average mean and variance underlying the two dashed CDFs and is the subjective CDF we use to calculate expected values in the MES data.

29. Fitting a beta distribution or using absolute deviations as an objective yields similar results.

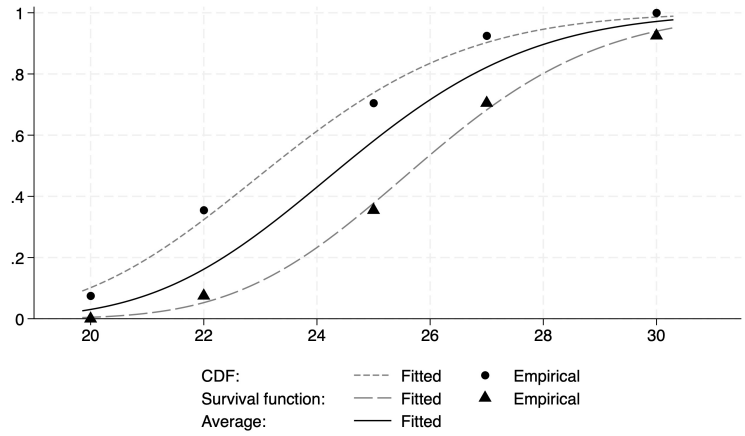
30. In the Cobb-Douglas specification, one only requires expectations rather than the subjective CDF, which could alternatively be estimated as a weighted sum across scenario values using the reported likelihoods as weights. These weighted sums are very similar to those of the fitted lognormals and we proceed with the latter so that we can analyze non-linear production function specifications using the same subjective expectations as those used to examine the Cobb-Douglas case. An alternative would be to rely on "bounds" defined by the CDF and survival functions.

Figure 2: Example of Fitting a Subjective CDF to Scenario-based Expectations

(a) Reported Scenario Values and Probabilities



(b) Subjective Distribution Functions



Note: This is an example of how we create subjective distribution functions from the reported scenario values and probabilities from the MES data. Panel (a) has the data where a respondent estimates there is, for example, a 35% probability that next year's sales will be £25,000, a 28% chance of £22,000, etc. The lower panel shows the two methods of translating these responses into a fitted lognormal distribution (the CDF vs. survivor function approaches are dashed lines). We calculate the solid line as a lognormal CDF using the average of the two means and variances. See text for details.

Table 5 summarises absolute deviations between firms' reported point values and those implied by the fitted lognormal CDFs. These differences are small across all variables for all industries implying the fitted lognormal distributions provide a good continuous approximation of the subjective distributions underlying firms' responses to the

discrete MES expectations questions and compares favorably with the fit obtained in related works such as Dominitz and Manski (1997).[31]

Table 5: Summary Statistics of Average Absolute Deviations Between Reported and Fitted Probabilities

	Mean	p25	p50	p75	N
	Electronics				
Turnover	0.028	0.019	0.026	0.035	472
Employment	0.026	0.018	0.025	0.034	472
	Retail				
Turnover	0.029	0.018	0.027	0.037	2084
Employment	0.028	0.018	0.025	0.035	2084
	Restaurants				
Turnover	0.032	0.020	0.029	0.039	462
Employment	0.032	0.020	0.029	0.038	462

Note: table contains means and quantile-group averages of the mean absolute deviations across firms' reported scenario likelihoods and those implied by the fitted lognormal distributions. The column titled 'Mean' contains the arithmetic mean calculated across the number of observations given in the column titled 'N'. Columns titled 'p25'/'p50'/'p75' contain the mean values calculated across the 50 observations that are closest to the 25th/50th/75th percentile respectively owing to data disclosure requirements.

Our method of obtaining subjective distributions from the MES responses yields a distribution of lognormal parameters and implied moments across MES respondents. Table 6 summarizes the medians of these parameters and a selection of moments across variables and samples.[32] The average σ parameter of the fitted subjective distributions indicates firms' uncertainty and shows that restaurants are slightly more uncertain on average about year-ahead turnover than firms in electronics or retail, and more substantially so regarding year-ahead employment. As well as greater within-firm uncertainty, the restaurants sector also exhibits greater across-firm variation in mean expectations. This can be seen in Figure 3, which plots kernel densities of firms' expected mean growth rate for turnover and employment.

31. See discussion in their section 3.5. Note that we confine analysis of firms' expectations here to turnover and employment as these are the key expectations variables required by the baseline NPR estimator. Similar results for firms' expectations of investment and intermediate inputs as reported in the 2017 MES are available from the authors on request.

32. Percentile statistics are prohibited from being exported from the secure server through which we access the MES and ABS data. Table 6 therefore reports 'fuzzy medians' of the firm-specific lognormal subjective distribution parameters, calculated as the mean value across the 50 observations closest to the median.

Table 6: Median Fitted Subjective Distribution Characteristics

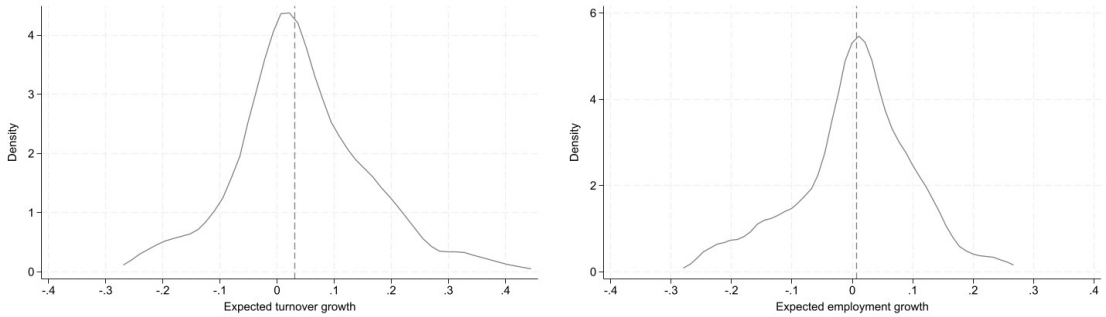
	μ	σ	Mean	S.D.	Median	IQR	N
	Electronics						
Sales Turnover £k	8.73	0.07	6196	3750	6181	511	472
Employment	3.89	0.04	48	29	48	3	472
	Retail						
Sales Turnover £k	8.97	0.06	7932	4823	7871	657	2084
Employment	3.74	0.05	42	26	42	3	2084
	Restaurants						
Sales Turnover £k	7.98	0.08	2848	1768	2811	321	462
Employment	4.27	0.09	67	41	67	8	462

Note: table contains mean values among the 50 observations closest to the median of the across-firm distribution of the parameter or moment indicated in the column title. Units for employment are for number of workers and for all other variables are thousands of pounds. The μ and σ columns give log-normal distribution parameters and therefore are interpreted as logged values of the variable given in the row title. The column 'N' indicates the number of observations for which we are able to fit a subjective distribution.

Figure 3: Expected growth rates

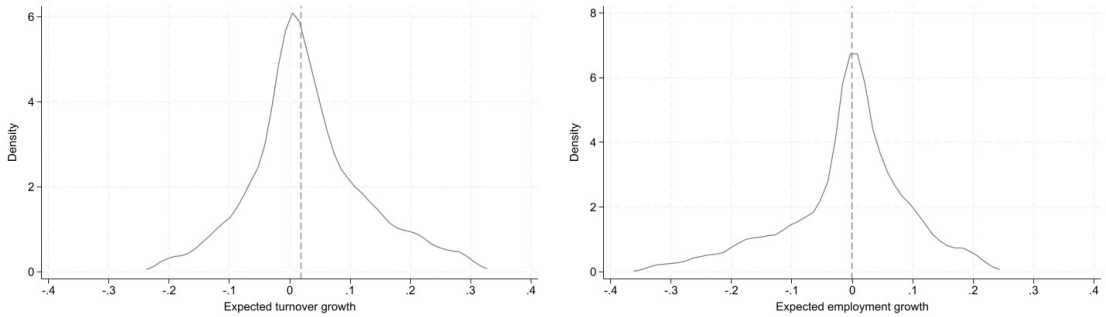
(a) Electronics: Turnover

(b) Electronics: Employment



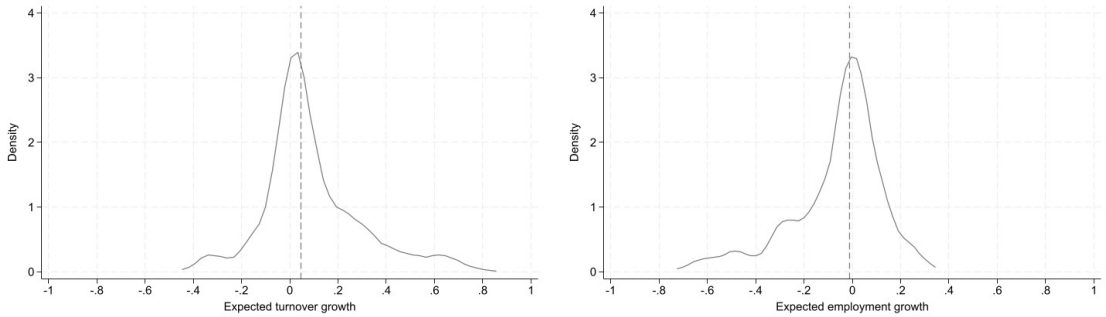
(c) Retail: Turnover

(d) Retail: Employment



(e) Restaurants: Turnover

(f) Restaurants: Employment



Note: Subfigures show kernel densities of $(\mathbb{E}_{it}[\ln(\mathrm{X})_{it+1}] - \ln(\mathrm{X})_{it})$ where X is either turnover or employment as indicated in the subfigure captions. Dashed vertical lines indicate medians. The top and bottom 5% of growth rates are included in the calculation of the medians but are not plotted in the figure.

A subsample of firms in the MES were also surveyed in the ABS the year following their MES response, which allows us to compare their subjective expectations to actual outcomes. Table 7 shows the log difference between firms' expectations and outcomes separately by variable and year. On average, firms' expectations are generally good, with

the forecast error being insignificantly different from zero. For employment the forecast error was zero for retail, 2% for electronics and -4% for restaurants. Electronics and retail industries also did reasonably well on turnover forecast (-1% and -5% respectively). Restaurants, however underestimated turnover by 18 log points (and 13 log points at the median).

Table 7: Log Deviation Between Expected Levels and Outcomes

	Mean	S.D.	p50	N obs.	N firms
	Electronics				
Turnover	-0.01	0.29	-0.04	221	205
Employment	0.02	0.18	0.02	221	205
	Retail				
Turnover	-0.05	0.25	-0.03	733	659
Employment	0.00	0.21	0.01	735	661
	Restaurants				
Turnover	-0.18	0.48	-0.13	170	150
Employment	-0.04	0.44	0.03	170	150

Note: The table contains summary statistics of the log deviation between the expected value and outcome of the variable denoted in the row title. Columns titled 'Mean' and 'S.D.' contain the mean and standard deviation respectively calculated across the number of firms in the column titled 'N'. The column titled 'p50' contains the mean value among the 50 observations closest to the median of the across-firm distribution denoted in the column title as data disclosure requirements prevent us from reporting the median value itself.

In theory, as we elaborate in Appendix D, one can leverage expectation errors of the type summarized in Table 7 as an additional source of identifying variation. This only holds, however, if firms' expectations about their future inputs are biased on average and if this bias is accompanied by bias in expected output of a magnitude that is consistent with the true production technology. We do not pursue this additional identification method in our empirical analysis as Table 7 shows firms' mean expectation errors are insignificantly different from zero across all industries we consider.[33]

33. When subjective beliefs and realisations are available, another possibility is to weight observations by the prediction quality (e.g. inversely proportional to forecast errors), leveraging observations with more accurate predictions. We are grateful to Moshe Buchinsky for this suggestion and leave it for future research.

5 Results

To evaluate the empirical performance of the NPR estimator, we implement a number of other popular production function estimators on the MES data separately for each of our focus industries. Tables 8 show estimates for each industry.[34]

Across all industries, NPR returns estimates that are plausible whereas the first-differenced and fixed-effect OLS specifications (contained in columns 'OLS FD' and 'OLS FE' respectively), return coefficients that are considerably lower than the OLS levels coefficients with statistically insignificant capital coefficients. More specific comparisons between the various estimators differ by industry. Within the electronics industry, NPR returns a labor coefficient of 0.9 and a capital coefficient of 0.21, which are both insignificantly different from the results obtained via linear OLS and ACF. Taking this comparison at face value, it suggests that both labor and capital are subject to adjustment costs of such magnitude that labor cannot respond to within-period changes in persistent productivity (absolving the linear OLS estimates of bias), and that there is a one-to-one mapping between persistent productivity and intermediate inputs conditional on labor and capital (i.e. the ACF monotonicity condition holds). OP and LP meanwhile, return significantly lower capital coefficients, which may be due to the functional dependence issue highlighted by ACF.

Differences between NPR and ACF are larger for the retail industry estimates (0.80 vs. 0.66 for the output elasticity of labor), yet still insignificant. NPR and ACF return capital coefficient estimates of 0.16 and 0.17 respectively, which are significantly lower than the linear OLS estimate of 0.25. This suggests a different form of bias in the OLS estimate than the typical expectation that endogenous responses of labor to productivity cause upward bias the OLS labor coefficient and downward bias in the capital coefficient. OP and LP again return unlikely estimates with both the labor and capital coefficients appearing attenuated toward zero.

Estimates from the restaurant sector show NPR and linear OLS are statistically indistinguishable with NPR returning labor and capital coefficients of 0.82 and 0.26 re-

34. All results discussed in this section are from a Cobb-Douglas gross output production function (i.e. one that takes turnover as the measure of output). Appendix D.1 contains an explanation of how one can construct expectations of (log) value added using expectations questions of the type contained in the MES for turnover and intermediates. We do not pursue estimation of a net production function however, as expected intermediates were only asked in MES 2017 leading to prohibitively small industry-specific samples. The number of firms in Table 8 is comparable across all the methods but the number of observations is about half as small for the NPR estimator than for the other methods. This is because the proxy variable approaches require using lagged values of capital and investment, whereas NPR uses a single cross-section of data from each MES wave. The number of observations is less than twice the number of firms because some firms are surveyed in both the 2017 and 2020 MES.

spectively (compared to 0.82 and 0.21 for OLS levels). OP and LP again exhibit an attenuated labor coefficient while, unlike the results in the other industries, the ACF estimate appear implausible with a labor coefficient of 1.06 and capital coefficient insignificant at 0.05.

Table 8: Production Function Coefficient Estimates

	NPR	ACF	OLS LVL	OLS FD	OLS FE	OP	LP
	A. Electronics						
β_l	0.90***	0.79***	0.86***	0.45***	0.47***	0.60***	0.62***
	(0.14)	(0.13)	(0.05)	(0.12)	(0.10)	(0.05)	(0.07)
β_k	0.21*	0.23***	0.23***	0.04	0.04	0.26*	0.34***
	(0.12)	(0.09)	(0.04)	(0.03)	(0.03)	(0.15)	(0.08)
$\beta_l + \beta_k$	1.11	1.02	1.09	0.49	0.51	0.86	0.96
CRS	0.00	0.88	0.00	0.00	0.00	0.36	0.66
N obs.	472	917	917	458	895	917	917
N firms	422	422	422	411	400	422	422
	B. Retail						
β_l	0.80***	0.66***	0.75***	0.46***	0.52***	0.63***	0.55***
	(0.11)	(0.11)	(0.03)	(0.09)	(0.07)	(0.02)	(0.04)
β_k	0.16***	0.17*	0.25***	-0.02	-0.02	0.01	0.16***
	(0.05)	(0.09)	(0.03)	(0.02)	(0.02)	(0.08)	(0.05)
$\beta_l + \beta_k$	0.96	0.83	1	0.43	0.49	0.64	0.71
CRS	0.00	0.15	0.8	0.00	0.00	0.00	0.00
N obs.	2084	4057	4057	2023	3967	4057	4057
N firms	1853	1853	1853	1807	1763	1853	1853
	C. Restaurants						
β_l	0.82***	1.06***	0.81***	0.63***	0.68***	0.69***	0.69***
	(0.07)	(0.15)	(0.06)	(0.11)	(0.10)	(0.05)	(0.05)
β_k	0.26***	0.05	0.21***	-0.01	0.02	0.09	0.25*
	(0.07)	(0.14)	(0.06)	(0.06)	(0.05)	(0.09)	(0.14)
$\beta_l + \beta_k$	1.08	1.1	1.01	0.62	0.7	0.78	0.94
CRS	0.00	0.18	0.33	0.00	0.00	0.03	0.71
N obs.	462	879	879	421	838	879	879
N firms	430	430	430	392	389	430	430

Note: dependent variable is log turnover. Parentheses contain standard errors. NPR standard errors calculated from 100 bootstrap replications. Column titles indicate estimation methods. 'OLS Lvls/FD/FE' denotes OLS estimation in levels/first-differences/levels with firm fixed effects. Row 'CRS' contains the p-value from a test that the labor and capital coefficient sum to 1. All specifications include survey year dummies. */**/*** denote significance at the 10/5/1 percent level respectively.

To what extent should we believe the NPR estimates instead of those obtained by the other methods? It is straightforward to discard the OLS First Difference (FD) and OLS Fixed effect (FE) results. The large attenuation observed across all three industries indicates that exploiting some of the cross sectional variation across firms is necessary to yield credible estimates (as emphasized *inter alia* by **strom_production_1999)**. OP and LP also appear to perform poorly in all industries, possibly due to the functional dependence concern highlighted by ACF. NPR and OLS levels are similar in Electronics and Restaurants, but in Retail the capital elasticity is 50% larger in OLS than NPR, and the latter has the advantage of being robust to simultaneity concerns. Note that there are more marked differences when a translog production function is estimated (see Appendix F.2 and below).

In the Electronics and Retail industries, NPR is also similar to ACF, but in Restaurants there are large differences. There are two points in support of the NPR results. First, the capital coefficient for ACF is very small (0.05) and insignificantly different from zero compared to a significant 0.26 for NPR. It is *ex ante* implausible to believe that capital does not matter at all in this sector.

The second point in support of NPR is that material input invertibility appears less likely to hold in the restaurant industry than in either electronics or retail. Evidence to this point comes from a subsample of firms in the MES that were surveyed by the ABS. Unlike the ABS, which asks firms for annual values retrospectively after the year is complete, the MES was dispatched to firms in Fall-Winter of the survey year. This means the 'current-year' values that firms were asked for (i.e. 2017 values for firms surveyed in MES 2017 and 2020 values for firms surveyed in 2020), were estimates made before the year was complete. Comparing levels of inputs and output that firms report in the MES survey year to the equivalent values observed in the ABS therefore gives an indication of how unpredictable inputs are, and thus how hard they are to optimize. Table 9 contains moments of the distribution of MES-ABS log differences separately by industry and variable. This shows turnover was relatively easy to forecast prior to year-end, with a median difference between MES reported values and the ABS equivalents of zero across all industries. For electronics and retail, employment was equally as predictable, it was only slightly less so in the restaurants sector with a median MES-ABS deviation of 3%. The predictability of material inputs, however, varies more markedly across industries. Focusing on the median to avoid the influence of outliers, the average MES-ABS difference for intermediates is small for electronics and retail firms at 2% and 1% respectively. For restaurants, however, it is much larger at 30 log points. Under the assumption of Leontief technology, the fact that employment was relatively unchanged

from the date at which restaurants were surveyed by the MES whereas intermediate inputs changed substantially implies intermediate inputs cannot be at their intended level for the year. As shown in Section 3, this optimization error in intermediate inputs undermines ACF, owing to violation of the monotonicity condition they require, and we therefore view Table 9 as further reason to doubt the ACF production function estimates for the restaurants industry in favour of the NPR estimates.[35]

Table 9: Moments of Within-Firm MES-ABS Log Differences

	Mean	S.D.	Median	N obs.	N firms
	Electronics				
Turnover	0.00	0.17	-0.00	284	252
Employment	0.02	0.17	0.00	284	252
Intermediates	-0.15	1.41	0.02	282	250
	Retail				
Turnover	0.01	0.15	-0.00	1152	1038
Employment	0.02	0.20	0.00	1151	1038
Intermediates	-0.26	1.37	0.01	1150	1038
	Restaurants				
Turnover	-0.00	0.34	-0.00	235	209
Employment	-0.03	0.38	0.03	236	209
Intermediates	0.19	1.03	**0.30**	236	209

Note: table shows summary statistics of $\left(\ln(X_{it}^{MES}) - \ln(X_{it}^{ABS})\right)$ where superscripts denote survey name. X is either turnover, employment or intermediates as indicated in the row title. t is either 2017 or 2020 according to which year firms were surveyed by the MES. Sample is restricted to firms in the MES analysis sample that are also observed in the wave of the ABS that records values for the same period as the MES survey year. The 'Median' column contains the mean values calculated across the 50 observations that are closest to the 50th percentile owing to data disclosure requirements.

While we have focused here on Cobb-Douglas production technology, the NPR estimator can also estimate parameters of a translog production function. Appendix F.2 describes the additional data preparation necessary for this alternate specification and presents estimates for our three industries along with alternative estimates obtained via various OLS estimators and the ACF method. Testing whether the additional nonlinear translog parameters are jointly zero, the NPR estimates fail to reject Cobb-Douglas technology for Electronics and Restaurant sectors, but does for Retail. However, the average partial derivatives are not significantly different from the Cobb-Douglas equivalents in

35. A third more technical point is to first note that the returns to scale implied by the ACF estimates are similar to those implied by the NPR estimates. This suggests that the ACF estimates may be affected by the global identification issue noted above in footnote 21.

all industries so we therefore confine attention to the Cobb-Douglas results in subsequent empirical analysis for parsimony.

5.1 Implications of Results: Productivity

The results in the previous subsection show that the NPR estimator recovers production function parameter estimates that are robust to simultaneity concerns and are either more credible than or similar to alternative standard estimators. Equipped with such estimates we now demonstrate their utility and examine trends in total factor productivity (TFP). To examine trends in TFP, we take the parameter estimates obtained using the MES data and calculate TFP for the entire ABS sample as:

$$\hat{a}_{it}^{m} = y_{it} - \hat{\beta}_{l}^{mj} l_{it} - \hat{\beta}_{k}^{mj} k_{it}, \tag{15}$$

where in this and subsequent equations, i denotes firm, j denotes industry, and t denotes year. $\hat{\beta}^{mj}$ represents production function coefficients obtained by implementing estimator m on MES data for industry j. For parsimony and in light of the results of the previous subsection, we focus on comparing the NPR estimates to those obtained by OLS levels and ACF because the other estimators did not generally yield credible estimates.

We calculate TFP according to equation (15) for all firms observed in our three industries in the ABS between 2010 and 2019 and de-mean TFP by year.[36] We relate firm performance to TFP using equations of the form:

$$y_{it} = \pi \hat{a}_{it}^{m} + \tau_t + \epsilon_{it}, \tag{16}$$

where y is one of several firm-level outcomes, τ is a year dummy, $\hat{a}^m$ the TFP estimate obtained using coefficients from estimator m and ϵ is a mean-zero disturbance. The outcomes we consider are exit, and employment growth over various time horizons:

$$\Delta_{t,t+s}(l) = l_{it+s} - l_{it+1},$$

where $s \in (2, 5)$.[37]

36. We focus on this period as the most recent decade around the MES survey years but the comparisons across estimators we highlight are qualitatively similar if one either uses a larger sample period of 2000-2019 or a narrower one of 2015-2019. We focus on the intermediate period to add credibility to the assumption of constant technology, which is imposed implicitly via our TFP calculations, while balancing statistical power due to sample size.

37. We take $t+1$ as the base year of our growth measures to avoid endogeneity with the TFP estimates $\hat{a}_{it}^{m}$, since these are a linear function of l_t.

Table 10 contains the estimated $\hat{\pi}$ parameters from equation (16) along with the mean values of the outcome variables and sample sizes. The first panel of the table presents results for the electronics industry. Higher TFP is associated with significantly lower firm exit and higher employment growth conditional on survival. The magnitude of association is similarly large across the NPR, OLS and ACF estimates of TFP, with a one standard-deviation increase in TFP reducing the probability of exit by around 1.6 percentage points, in comparison to an overall exit rate of 1%, and increasing one- and four-year employment growth by 0.01 and 0.04 log points in comparison to mean growth rates of 0.01 and 0.025 respectively. Similar associations are observed within the retail industry, although the positive association between TFP and five-year employment growth is stronger for the NPR and ACF TFP estimates than for OLS-estimated TFP.

The third panel of Table 10 shows the restaurants industry exhibits greater variation across associations between firm performance and the various TFP estimates. Both NPR and ACF estimates of TFP are negatively associated with firm exit, which is somewhat surprising given the latter are somewhat implausible. This is in contrast to the OLS TFP estimate, which has no significant association with firm exit. All three TFP estimates are significantly positively associated with two-year employment growth but the strength of association varies from 0.012 with the OLS TFP estimate to 0.021 with the NPR TFP estimate. Despite being small in absolute terms, this difference is considerable and corresponds to 22% and 39% of the mean growth rate of 0.054. Point estimates of the association between TFP and employment growth over a four-year horizon indicate the association is again strongest for the NPR TFP estimate, although the smaller sample size means none of the point estimates or their differences are significant. Viewed together, these results suggest that failing to adequately account for input endogoeneity risks underestimating the degree of dynamic reallocation over a mid-range (e.g. four-year) horizon.

Table 10: Association Between TFP and Firm Performance

	NPR	ACF	OLS	Mean	N
	Electronics				
Exit	-0.016***	-0.017***	-0.016***	0.01	5362
	(0.004)	(0.004)	(0.004)		
$\Delta_{t+1,t+2}(l)$	0.010**	0.008*	0.009**	0.01	2201
	(0.005)	(0.005)	(0.005)		
$\Delta_{t+1,t+5}(l)$	0.044**	0.03	0.041**	0.025	1398
	(0.021)	(0.020)	(0.020)		
	Retail				
Exit	-0.014***	-0.014***	-0.014***	0.024	57630
	(0.001)	(0.001)	(0.001)		
$\Delta_{t+1,t+2}(l)$	0.003***	0.003***	0.003***	0.026	13485
	(0.001)	(0.001)	(0.001)		
$\Delta_{t+1,t+5}(l)$	0.007*	0.005	0.008**	0.095	8331
	(0.004)	(0.004)	(0.004)		
	Restaurants				
Exit	-0.007***	-0.003	-0.011***	0.038	12513
	(0.002)	(0.002)	(0.002)		
$\Delta_{t+1,t+2}(l)$	0.021***	0.012*	0.016**	0.054	2643
	(0.007)	(0.007)	(0.007)		
$\Delta_{t+1,t+5}(l)$	0.036	0.004	0.005	0.225	1581
	(0.031)	(0.028)	(0.030)		

Note: table shows OLS estimates of the parameter on TFP from regressions that relate the outcome given in the row title to TFP implied by the production function estimator denoted by the column title. 'Exit' is a dummy variable that takes the value of 1 in period t if the firm is recorded as dead in administrative data in period $t+1$ and zero otherwise. $\Delta_{t,t+s}(l)$ are log differences in labor between period $t+1$ and $t+s$. Data is taken from the ABS between 2010 and 2019 inclusive. $^{*}/^{**}/^{***}$ denote significance at the 10/5/1 percent level respectively.

6 Conclusion

In this paper we have proposed a new production function estimation methodology that leverages data on firms' observable expectations, data which are becoming increasingly available across a range of countries. We show that such information enables one to relax the strong assumptions of firm input choice (strict monotonicity and invertibility with respect to a single scalar unobservable productivity term), that underpin currently-used proxy variable approaches in the Olley and Pakes (1996) tradition. Moreover, our method can be implemented on a single cross section which has great attractions over these techniques (which require two or more periods), and dynamic panel methods such

as those of Blundell and Bond (2000), which assume linear productivity dynamics and require three or four consecutive time period observations per firm.

We present Monte Carlo simulations close to those in Ackerberg et al. (2015) featuring forward-looking firms with heterogeneous quadratic adjustment costs in capital. We show that our proposed NPR estimator is robust to optimization error in firm inputs choices whereas other methods are biased when firms make optimization errors in their material input choices. We also demonstrate the NPR estimator can be extended to accommodate certain forms of bias in firms' expectations.

Implementing our proposed NPR estimator on UK data, we show it recovers parameter estimates that are either comparable to or more credible than those recovered by conventionally-used estimators. We use the various production function estimates to calculate TFP residuals and relate these to measures of firm performance. TFP is negatively associated with firm exit across all industries and production function estimators we consider while TFP estimates obtained using our proposed production function estimator are more positively associated with employment growth, especially over a mid-range horizon. This suggests that the extent of dynamic reallocation is particularly sensitive to adequately accounting for input endogeneity during production function estimation.

References

Ackerberg, Kevin Caves, and Garth Frazer. 2015. "Identification Properties of Recent Production Function Estimators." *Econometrica* 83 (6): 2411–2451.

Ackerberg, C. Lanier Benkard, Steven Berry, and Ariel Pakes. 2007. "Chapter 63 Econometric Tools for Analyzing Market Outcomes." In *Handbook of Econometrics,* 6:4171–4276. Elsevier.

Altig, David, Jose Maria Barrero, Nicholas Bloom, Steven J. Davis, Brent Meyer, and Nicholas Parker. 2022. "Surveying business uncertainty." Annals Issue: Subjective Expectations and Probabilities in Economics, *Journal of Econometrics* 231 (1): 282–303.

Anderson, T. W., and Cheng Hsiao. 1981. "Estimation of Dynamic Models with Error Components." *Journal of the American Statistical Association* 76, no. 375 (September): 598–606.

Arellano, Manuel, Orazio Attanasio, Samuel Crossman, and Victor Sacibrian. 2024. *Estimating Flexible Income Processes from Subjective Expectations Data: Evidence from India and Colombia.* w32922. National Bureau of Economic Research.

Attanasio, Orazio, Flávio Cunha, and Pamela Jervis. 2024. "Subjective Beliefs: Their Measurement and Role." *Working Paper.*

Barrero, Jose. 2022. "The micro and macro of managerial beliefs." *Journal of Financial Economics* 143 (2): 640–667.

Bierens, Herman J. 1983. "Uniform Consistency of Kernel Estimators of a Regression Function Under Generalized Conditions." *Journal of the American Statistical Association* 78 (383): 699–707.

Bloom, Nicholas, Erik Brynjolfsson, Lucia Foster, Ron Jarmin, Megha Patnaik, Itay Saporta-Eksten, and John Van Reenen. 2019. "What Drives Differences in Management Practices?" *American Economic Review* 109, no. 5 (May): 1648–1683.

Bloom, Nicholas, Philip Bunn, Paul Mizen, Pawel Smietanka, Gregory Thwaites, and Garry Young. 2017. *Tracking the Views of British Businesses: Evidence Form the Decision Maker Panel,* 3004466, Rochester, NY, June 16, 2017.

Bloom, Nicholas, Mihai Codreanu, and Robert Fletcher. 2025. *Rationalizing Firm Forecasts.* Working Paper, w33384. National Bureau of Economic Research.

Bloom, Nicholas, Takafumi Kawakubo, Charlotte Meng, Paul Mizen, Rebecca Riley, Tatsuro Senga, and John Van Reenen. 2026. "Do Well Managed Firms Make Better Forecasts?" *Review of Economics and Statistics.*

Blundell, Richard, and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics* 87, no. 1 (November): 115–143.

———. 2000. "GMM Estimation with persistent panel data: an application to production functions." *Econometric Reviews* 19, no. 3 (January 1, 2000): 321–340.

Bond, Stephen, and M Söderbom. 2005. *Adjustment costs and the identification of Cobb Douglas production functions.* Working Paper Series. Series: Working Paper Series. February 9, 2005.

Born, Benjamin, Zeno Enders, Gernot J. Müller, and Knut Niemann. 2022. "Firm expectations about production and prices: Facts, determinants, and effects." Chap. 12 in *"Handbook of Economic Expectations",* edited by "Rüdiger Bachmann, Giorgio Topa, and Wilbert van der Klaauw", 355–384. "Amsterdam": "Elsevier".

Buffington, Catherine, Lucia Foster, Ron S. Jarmin, and Scott Ohlmacher. 2016. *The Management and Organizational Practices Survey (MOPS): An Overview,* 2793858, Rochester, NY, June 1, 2016.

Cobb, Charles W., and Paul H. Douglas. 1928. "A Theory of Production." *The American Economic Review* 18 (1): 139–165.

Collard-Wexler, Allan, and Jan De Loecker. 2021. "Productivity and Capital Measurement Error." *in Progress.*

Cunha, Flavio, and James J. Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *The Journal of Human Resources* 43 (4): 738–782.

De Loecker, Jan. 2011. "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity." *Econometrica* 79 (5): 1407–1451.

De Loecker, Jan, Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik. 2016. "Prices, Markups, and Trade Reform." *Econometrica* 84 (2): 445–510.

De Loecker, Jan, and Chad Syverson. 2021. "An industrial organization perspective on productivity." In *Handbook of Industrial Organization,* 4:141–223. Elsevier.

De Loecker, Jan, and Frederic Warzynski. 2012. "Markups and Firm-Level Export Status." *American Economic Review* 102, no. 6 (May): 2437–2471.

De Roux, Nicolás, Marcela Eslava, Santiago Franco, and Eric Verhoogen. 2021. *Estimating Production Functions in Differentiated-Product Industries with Quantity Information and External Instruments.* w28323. Cambridge, MA: National Bureau of Economic Research, January.

Demirer, Mert. 2022. "Production Function Estimation with Factor-Augmenting Technology: An Application to Markups."

Dominitz, Jeff, and Charles F. Manski. 1997. "Using Expectations Data to Study Subjective Income Expectations." *Journal of the American Statistical Association* 92, no. 439 (September): 855–867.

Dominitz, Jeff, and Robert P. Sherman. 2005. "Some Convergence Theory for Iterative Procedures with an Application to Semiparametric Estimation." *Econometric Theory* 21 (4): 838–863.

Doraszelski, Ulrich, and Jordi Jaumandreu. 2013. "R&D and Productivity: Estimating Endogenous Productivity." *The Review of Economic Studies* 80 (4): 1338–1383.

———. 2018. "Measuring the Bias of Technological Change." *Journal of Political Economy* 126, no. 3 (June): 1027–1084.

Dovern, Jonas, Lena Sophia Müller, and Klaus Wohlrabe. 2023. "Local information and firm expectations about aggregates." *Journal of Monetary Economics* 138:1–13.

Eslava, Marcela, John Haltiwanger, and Nicoloas Urdaneta. 2026. "The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand and Wedges." *Review of Economic Studies* 91 (1): 259–300.

Evdokimov, Kirill S., and Andrei Zeleneev. 2025. *Simple estimation of semiparametric models with measurement errors.* CEMMAP Working Paper, CWP02/25. Centre for Microdata Methods and Practice.

Friedman, Jerome H., and Werner Stuetzle. 1981. "Projection Pursuit Regression." *Journal of the American Statistical Association* 76, no. 376 (December): 817–823.

Gandhi, Amit, Salvador Navarro, and David A. Rivers. 2020. "On the Identification of Gross Output Production Functions." Publisher: The University of Chicago Press, *Journal of Political Economy* 128, no. 8 (August): 2973–3016.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer. 2016. "Expectations and Investment." *NBER Macroeconomics Annual* 30, no. 1 (January): 379–431.

Griliches, Zvi, and Jacques Mairesse. 1999. "Production Functions: The Search for Identification." In *Econometrics and Economic Theory in the 20th Century,* edited by Steinar Strom, 169–203. Cambridge: Cambridge University Press.

Hall, Robert E. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy* 96, no. 5 (October): 921–947.

Hastie, Trevor, and Robert Tibshirani. 1986. "Generalized Additive Models." *Statistical Science* 1, no. 3 (August 1, 1986).

Horowitz, Joel L., and Enno Mammen. 2004. "Nonparametric Estimation of an Additive Model with a Link Function." *The Annals of Statistics* 32 (6): 2412–2443.

Ichimura, Hidehiko, and Petra E. Todd. 2007. "Chapter 74 Implementing Nonparametric and Semiparametric Estimators." In *Handbook of Econometrics,* 6:5369–5468. Elsevier.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money.* Macmillan,London.

Levinsohn, James, and Amil Petrin. 2003. "Estimating Production Functions Using Inputs to Control for Unobservables." *The Review of Economic Studies* 70 (2): 317–341.

Manski, Charles F. 2004. "Measuring Expectations." *Econometrica* 72 (5): 1329–1376.

Marschak, Jacob, and William H. Andrews. 1944. "Random Simultaneous Equations and the Theory of Production." *Econometrica* 12, no. 3 (July): 143.

Mundlak, Yair. 1961. "Empirical Production Function Free of Management Bias." *Journal of Farm Economics* 43, no. 1 (February): 44.

Nadarajah, Saralees, and Samuel Kotz. 2008. "Exact Distribution of the Max/Min of Two Gaussian Random Variables." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 16, no. 2 (February): 210–212.

Newey, Whitney K., and Daniel McFadden. 1994. "Chapter 36 Large sample estimation and hypothesis testing," 4:2111–2245. Handbook of Econometrics. Elsevier.

Office For National Statistics. 2022. *MES Management and Expectations Survey, 2016-2020: Secure Access.* Type: dataset.

———. 2023. *ABS Annual Business Survey, 2005-2021: Secure Access.* V. 17th Edition.

Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." Publisher: [Wiley, Econometric Society], *Econometrica* 64 (6): 1263–1297.

Orr, Scott. 2022. "Within-Firm Productivity Dispersion: Estimates and Implications." *Journal of Political Economy* 130, no. 11 (November 1, 2022): 2771–2828.

Pakes, Ariel. 1994. "Dynamic structural models, problems and prospects: mixed continuous discrete controls and market interactions." In *Advances in Econometrics: Sixth World Congress,* edited by Christopher A.Editor Sims, 171–274. Econometric Society Monographs. Cambridge University Press.

Pastorello, Sergio, Valentin Patilea, and Eric Renault. 2003. "Iterative and Recursive Estimation in Structural Nonadaptive Models." *Journal of Business and Economic Statistics* 21 (4): 449–482.

Pesaran, M. Hashem, and Martin Weale. 2006. "Chapter 14 Survey Expectations," edited by G. Elliott, C.W.J. Granger, and A. Timmermann, 1:715–776. Handbook of Economic Forecasting. Elsevier.

Pistaferri, Luigi. 2001. "Superior Information, Income Shocks, And The Permanent Income Hypothesis." *The Review of Economics and Statistics* 83 (3): 465–476.

Pya, Natalya, and Simon N. Wood. 2015. "Shape constrained additive models." *Statistics and Computing* 25, no. 3 (May): 543–559.

Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56, no. 4 (July): 931.

Rovigatti, Gabriele, and Vincenzo Mollisi. 2018. "Theory and practice of total-factor productivity estimation: The control function approach using Stata." *Stata Journal* 18 (3): 618–662.

Solow, Robert M. 1957. "Technical Change and the Aggregate Production Function." *The Review of Economics and Statistics* 39, no. 3 (August): 312.

Syverson, Chad. 2001. "Market Structure and Productivity." *Ph.D. Dissertation, University of Maryland.*

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113, no. 485 (February 1, 2003): F3–F33.

Valmari, Nelli. 2023. "Estimating Production Functions of Multiproduct Firms." *Review of Economic Studies* 90 (6): 3315–3342.

Van Biesebroeck, Johannes. 2007. "Robustness of Productivity Estimates." *The Journal of Industrial Economics* 55 (3): 529–569.

Wang, Xiao, Jinglai Shen, and David Ruppert. 2011. "On the asymptotics of penalized spline smoothing." *Electronic Journal of Statistics* 5 (none): 1–17.

Xiao, Luo. 2019. "Asymptotic theory of penalized splines." *Electronic Journal of Statistics* 13 (1): 747–794.

Zellner, A., J. Kmenta, and J. Dreze. 1966. "Specification and Estimation of Cobb-Douglas Production Function Models." *Econometrica* 34, no. 4 (October): 784.

Appendices

A Expectation formation processes

As noted in the main text, we mirror the notation used in Pesaran and Weale 2006 and use the operator $\mathbb{E}_{it}[\cdot|\Omega_{it}]$ to denote a firm's expectation given its information set when applied to output y_{it} and inputs (e.g., l_{it}): $\mathbb{E}_{it}[\mathbf{x}_{it+1}|\Omega_{it}] = \int \mathbf{x} h_{it}(\mathbf{x}|\Omega_{it})d\mathbf{x}$, where $\mathbf{x}$ is a vector and $h_{it}(\cdot)$ represents firm i's distribution for it in period t given their information Ω_{it}. As long as they satisfy the identification conditions in Theorem 1 and Corollary 1 (see below), we do not take a stand on the expectation formation process for output or inputs and treat those as data. One may nonetheless envision different scenarios on the expectation process for productivity. Pesaran and Weale discuss several expectation formation protocols (see their Section 2) and we borrow from their discussion here.

Rational Expectations: This case corresponds to: $\mathbb{E}_{it}[\omega_{it}|\Omega_{it-1}] = \mathbb{E}[\omega_{it}|\Omega_{it-1}]$, which is used to obtain equation (2) in the main text:

$$\begin{aligned}\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] &= \int f(k_{it+1}, l_{it+1}; \beta)dF_{it}(l_{it+1}) + \mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] + \mathbb{E}_{it}[\epsilon_{it+1}|\Omega_{it}] \\ &= \int f(k_{it+1}, l_{it+1}; \beta)dF_{it}(l_{it+1}) + g(\omega_{it}),\end{aligned}$$

where $F_{it}(l_{it+1})$ represents firm i's subjective probability distribution over their next-period labor input given its information Ω_{it} and the second equality follows from the assumptions that $\mathbb{E}_{it}[\epsilon_{it}|\Omega_{it-1}] = \mathbb{E}[\epsilon_{it}|\Omega_{it-1}]$ and $\mathbb{E}_{it}[\xi_{it}|\Omega_{it-1}] = \mathbb{E}[\xi_{it}|\Omega_{it-1}]$ are equal to zero.

Static Expectations: This is the simplest form of expectation formation protocol dating back to Keynes (1936) and amounts to $\mathbb{E}_{it}[\mathbf{x}_{it+1}|\Omega_{it}] = \mathbf{x}_{it}$. One can envision then a process by which $\mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] = \tilde{g}(\omega_{it})$ where $\tilde{g}(\cdot)$ is potentially different from $g(\cdot)$. In this case,

$$\begin{aligned}\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] &= \int f(k_{it+1}, l_{it+1}; \beta)dF_{it}(l_{it+1}) + \mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] + \mathbb{E}_{it}[\epsilon_{it+1}|\Omega_{it}] \\ &= \int f(k_{it+1}, l_{it+1}; \beta)dF_{it}(l_{it+1}) + \tilde{g}(\omega_{it}),\end{aligned}$$

where once again we assume that $\mathbb{E}_{it}[\epsilon_{it}|\Omega_{it-1}] = 0$ which seems reasonable since by definition ϵ_{it} is either measurement error or not in the information set for the firm. In this

case, as long as $\tilde{g}(\cdot)$ is invertible, $\Psi(\cdot)$ is not $g^{-1}(\cdot)$ but $\tilde{g}^{-1}(\cdot)$.

Return to normality models: Pesaran and Weale 2006 discuss this case in terms of the average or consensus expectation. For an individual firm, expectations are presumed to follow: $\mathbb{E}_{it}[\mathbf{x}_{it+1}|\Omega_{it}] = (\mathbf{I} - \mathbf{\Lambda}_i)\mathbf{x}_{it} + \mathbf{\Lambda}_i\mathbf{x}^*_{it}$ where $\mathbf{x}^*_{it}$ represents the 'normal' or 'long run equilibrium' level of $\mathbf{x}_{it}$. Here, $\mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] = (1-\lambda)\omega_{it} + \lambda\omega^*_i$ where we impose homogeneity in λ across firms and a time-invariant 'normal' or 'long run equilibrium' level. Then,

$$\begin{aligned}\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] &= \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) + \mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] + \mathbb{E}_{it}[\epsilon_{it+1}|\Omega_{it}] \\ &= \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) + (1-\lambda)\omega_{it} + \lambda\omega^*_i,\end{aligned}$$

where once again we assume that $\mathbb{E}_{it}[\epsilon_{it}|\Omega_{it-1}] = 0$. In this case, one ends up with $\Psi(x) = x/(1-\lambda)$ and $-\lambda\omega^*_i/(1-\lambda)$ in equation (9), for example. If $-\lambda\omega^*_i/(1-\lambda)$ is homogeneous across firms within an industry then an industry-specific estimate as we perform here will be adequate. If it is a firm "fixed effect" then we need a panel as discussed below in Appendix D (Case 2).

Adaptive Expectations: This case amounts to $\mathbb{E}_{it}[\mathbf{x}_{it+1}|\Omega_{it}] = \mathbb{E}_{it-1}[\mathbf{x}_{it}|\Omega_{it-1}] + \mathbf{\Gamma}_i(\mathbf{x}_{it} - \mathbb{E}_{it-1}[\mathbf{x}_{it}|\Omega_{it-1}])$ (once again specialising this to the individual firm) and the expectations can also be represented as an infinite distributed lag process. For a unidimensional case considering ω_{it}, this means that $\mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] = \gamma_i\omega_{it} + (1-\gamma_i)\mathbb{E}_{it-1}[\omega_{it}|\Omega_{it-1}] = \gamma_i[\omega_{it} + (1-\gamma_i)\omega_{it-1} + (1-\gamma_i)^2\omega_{it-2} + \dots]$. Assuming that $\gamma_i = \gamma$ for any i, we then have that

$$\begin{aligned}\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] &= \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) + \mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}] + \mathbb{E}_{it}[\epsilon_{it+1}|\Omega_{it}] \\ &= \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) + \gamma[\omega_{it} + (1-\gamma)\omega_{it-1} + (1-\gamma)^2\omega_{it-2} + \dots].\end{aligned}$$

Whereas this breaks the Markovianity assumption we rely on, we can adapt usual protocols in the estimation of geometric (Koyck) models. Namely, notice that

$$\begin{aligned}\mathbb{E}_{it-1}[y_{it}|\Omega_{it-1}] &= \int f(k_{it}, l_{it}; \beta) dF_{it}(l_{it}) + \mathbb{E}_{it-1}[\omega_{it}|\Omega_{it-1}] + \mathbb{E}_{it-1}[\epsilon_{it}|\Omega_{it-1}] \\ &= \int f(k_{it}, l_{it}; \beta) dF_{it-1}(l_{it}) + \gamma[\omega_{it-1} + (1-\gamma)\omega_{it-2} + (1-\gamma)^2\omega_{it-3} + \dots].\end{aligned}$$

This implies that

$$\gamma[(1-\gamma)\omega_{it-1}+(1-\gamma)^2\omega_{it-2}+\ldots] = (1-\gamma)\{\mathbb{E}_{it-1}[y_{it}|\Omega_{it-1}] - \int f(k_{it},l_{it};\beta)dF_{it-1}(l_{it})\}.$$

Consequently, we obtain that

$$\begin{aligned}\mathbb{E}_{it}[y_{it+1}|\Omega_{it}] &= \int f(k_{it+1},l_{it+1};\beta)dF_{it}(l_{it+1}) \\ &\quad +(1-\gamma)\{\mathbb{E}_{it-1}[y_{it}|\Omega_{it-1}] - \int f(k_{it},l_{it};\beta)dF_{it-1}(l_{it})\} + \gamma\omega_{it}.\end{aligned}$$

Thus, we get that $\Psi(x) = x/\gamma$ and it takes as argument the original expression minus $(1-\gamma)\{\mathbb{E}_{it-1}[y_{it}|\Omega_{it-1}] - \int f(k_{it},l_{it};\beta)dF_{it-1}(l_{it})\}$ in equation (9), for example. Notice that this would require that observations on expectations be collected in t and $t-1$.

Conclusions on Expectations Processes: The discussion here shows that NPR (or extensions of it) can handle a wide range of canonical expectations processes. Although we have developed our formal proofs for NPR under rational expectations, it encompasses a wider class of firm forecasting protocols.

B Convergence of the NPR estimator

The demonstration here follows Theorem 2 in Dominitz and Sherman 2005 (DS). We adopt the notation in Theorem 1 and Corollary 1 in our paper, suppose that $f(\cdot)$ is linear and focus on a cross-section for simplicity:

$$y_i = x_i^\top\beta_0 + h_0(z_i) + \epsilon_i = x_i^\top\beta_0 + \Psi_0(z_{1i} - z_{2i}^\top\beta_0) + \epsilon_i, \tag{17}$$

where $x_i \in \mathbb{R}^k$ (e.g., $(\ln k_i, \ln l_i) \in \mathbb{R}^2$), $z_i = [z_{i1}\ z_{i2}^\top]^\top$ with $z_{i1} \in \mathbb{R}$ (i.e., expected log-production next period) and $z_{i2} \in \mathbb{R}^k$ (i.e., log-capital and expected log-labor next period). In what follows, we let $\mathcal{N} \subset \mathbb{R}^k$ denote the closure of an open convex neighbourhood of β_0.

We define two objects as in DS. For $\phi \in \mathbb{R}^k$, let

$$M_n(\phi) = \arg\min_\beta \sum_{i=1}^n (y_i(\phi) - x_i^\top\beta)^2 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}(\phi) \in \mathbb{R}^k, \tag{18}$$

where $y_i(\phi) \equiv y_i - \Psi_0(z_{1i} - \phi^\top z_{2i}) - \epsilon_i$, $\mathbf{X}_{n\times k} = [x_1\ \cdots\ x_n]^\top$ and $\mathbf{Y}(\phi)_{n\times 1} = [y_1(\phi)\ldots y_n(\phi)]^\top$.

Note that $M_n(\beta_0) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}(\beta_0) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\beta_0 = \beta_0$. We later also use the matrix $\mathbf{Z}_{2\ (n\times k)} = [z_{21}\ \cdots\ z_{2n}]^\top$. Furthermore, let

$$\widehat{M}_n(\phi) = \arg\min_{\beta} \sum_{i=1}^{n} (\hat{y}_i(\phi) - x_i^\top\beta)^2 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\widehat{\mathbf{Y}}(\phi) \in \mathbb{R}^k, \tag{19}$$

where $\hat{y}_i(\phi) \equiv y_i - \hat{\Psi}(z_{1i} - \phi^\top z_{2i})$ and $\widehat{\mathbf{Y}}(\phi)_{n\times 1} = [\hat{y}_1(\phi)\ \cdots\ \hat{y}_n(\phi)]^\top$. Here, $\hat{\Psi}(\cdot)$ is an estimate for $\Psi_0(\cdot)$, which is constructed using splines in our application. $M_n(\cdot)$ and $\widehat{M}_n(\cdot)$ relate to the steps in our algorithm. As noted in DS, for $i \geq 1$, $\hat{\beta}_i = \widehat{M}_n(\hat{\beta}_{i-1})$ and $\beta_i = M_n(\beta_{i-1})$ correspond to sample and population iterates in the iterative protocol.

Following the notation in DS, we also define ∇_ϕ as the differential operator with respect to ϕ. Consequently, $\nabla_\phi M_n(\phi)_{k\times k}$ is the matrix of first-order derivatives for $M_n(\phi)$ and similarly for $\nabla_\phi \widehat{M}_n(\phi)_{k\times k}$. In particular, notice that $\nabla_\phi M_n(\phi)_{k\times k} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\nabla_\phi\mathbf{Y}(\phi)$ and $\nabla_\phi \widehat{M}_n(\phi)_{k\times k} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\nabla_\phi\widehat{\mathbf{Y}}(\phi)$, where $\nabla_\phi\mathbf{Y}(\phi)_{n\times k}$ and $\nabla_\phi\widehat{\mathbf{Y}}(\phi)_{n\times k}$ are derivative matrices for $\mathbf{Y}(\phi)$ and $\widehat{\mathbf{Y}}(\phi)$, respectively.

To demonstrate convergence for our protocol, we use the following *sufficient* conditions:

Assumption 1 (DGP and Sampling)**.** *The data is generated according to Equation (17), where*

i. $x_i, z_i \perp\!\!\!\perp \epsilon_i$*;*

ii. $\mathbb{E}[x_i x_i^\top]$ *is non-singular and* $\sum_{i=1}^n |x_{ik_1} z_{2i,k_2} z_{2i,k_3}|/n = O_p(1)$*;*

iii. $(y_i, x_i, z_i)_{i=1}^n$ *is i.i.d.;*

iv. β_0 *belongs to a compact subset of* $\mathbb{R}^k$*; and*

v. $\Psi_0(\cdot) \in C_2$ *with bounded second-order derivative.*

In the above, x_{ik_1} is the k_1th entry in the vector x_i and z_{2i,k_2} and z_{2i,k_3} are the k_2th and k_3th entries in the vector z_{2i} where $1 \leq k_1, k_2, k_3 \leq k$. Assumption 1 (i), (ii) and (iii) are conventional. We suppose above that $\Psi_0(\cdot)$ is sufficiently smooth (i.e., continuously differentiable) with a bounded derivative in Assumption 1 (iv). If the law of motion for ω is a linear autoregressive process with coefficient ρ_0, $\Psi_0 : x \mapsto x/\rho_0$ and the assumption holds as long as ρ_0 is finite. In addition, we assume:

Assumption 2 (Estimation Protocol)**.** *The estimation is such that*

i. $\mathbb{P}(\sigma_{max}(\nabla_\phi M_n(\beta_0)) < 1) \to 1$*; and*

ii. $\hat{\Psi}(\cdot)$ *converges to* $\Psi_0(\cdot)$ *uniformly in probability.*

Assumption 2 (i) requires that the maximum singular value for $\nabla_\phi M_n(\beta_0)$ be strictly less than 1 with probability approaching one as the sample since increases. If the law of motion for ω is a linear autoregressive process with coefficient ρ_0, the assumption is satisfied if $\sigma_{\max}(\nabla_\phi M_n(\beta_0)) = \sigma_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}_2)/\rho_0 < 1$ with probability approaching 1 when $n \to \infty$. Since the maximum singular value of a matrix is a continuous function of the matrix, by the Continuous Mapping Theorem the assumption is satisfied as long as $\sigma_{\max}(\mathbb{E}[x_i x_i^\top]^{-1}\mathbb{E}[x_i z_i^\top])/\rho_0 < 1$. For more general laws of motion, one can in principle enforce this latter assumption on the estimation by constraining $\sup|\widehat{\Psi}'(\cdot)| < \sigma_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^{-1}\sigma_{\max}(\mathbf{Z}_2)^{-1}$.[38]

Assumption 2 (ii) in turn requires that $\hat{\Psi}(\cdot)$ converge to $\Psi_0(\cdot)$ uniformly in probability. Conditions for uniform consistency for (non-penalized) splines can be found, for example, in Horowitz and Mammen 2004 (see Theorem 1c) and for penalized splines in Wang et al. 2011 or Xiao 2019 (see Section 6). For kernel methods, results can be found, for example, in Bierens 1983.

As previously defined, let $\hat{\beta}_i = \widehat{M}_n(\hat{\beta}_{i-1})$ denote the sample iterate in the iterative protocol. As in DS, we allow i to depend on n and denote this by the function $i(n)$ mapping $n \in \mathbb{Z}^+$ into $i \in \mathbb{Z}^+$. Below, we use $\|\cdot\|$ to denote the Euclidean norm. We can then state the following convergence result:

Proposition 1. *Suppose* $i(n) \to \infty$ *as* $n \to \infty$. *If Assumptions 1 and 2 hold, then* $\|\hat{\beta}_{i(n)} - \beta_0\| = o_p(1)$ *as* $n \to \infty$.

This result follows from Theorem 2 in DS. This theorem relies on three conditions. The first condition requires that $M_n(\phi)$ be an Asymptotic Contraction Mapping (ACM), which we define below, and the third condition requires that $\|\widehat{M}_n(\phi) - M_n(\phi)\|$ converge to zero uniformly in probability. The second condition is implied by the first one if $i(n) \to \infty$ as $n \to \infty$. In order to demonstrate convergence, we thus establish two lemmas. Lemma 1 obtains that $M_n(\phi)$ is an Asymptotic Contraction Mapping (ACM) and Lemma 2 obtains that $\|\widehat{M}_n(\phi) - M_n(\phi)\|$ converges to zero uniformly in probability. Proposition thus follows.

The following notation is used below. For a matrix $A_{k\times k}$, let $\|A\|_2$ denote the matrix norm induced by the Euclidean norm. It coincides with the Euclidean norm for vectors

38. Notice that $\nabla_\phi\mathbf{Y}(\phi)$ defined in the proof for Lemma 1 is $\text{diag}(\Psi'(\mathbf{Z}))\mathbf{Z}_2$ where $\Psi'(\mathbf{Z})_{n\times 1} = [\Psi'(z_{21}) \cdots \Psi'(z_{2n})]^\top$. Then, $\sigma_{\max}(\nabla_\phi M_n(\beta_0)) = \sigma_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{diag}(\Psi'(\mathbf{Z}))\mathbf{Z}_2)$. By Theorem 3.3.4 in Horn and Johnson (Topics in Matrix Analysis), $\sigma_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{diag}(\Psi'(\mathbf{Z}))\mathbf{Z}_2) \leq \sigma_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\sigma_{\max}(\text{diag}(\Psi'(\mathbf{Z})))\sigma_{\max}(\mathbf{Z}_2) = \max_i|\Psi'(z_{2i})| \times \sigma_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\sigma_{\max}(\mathbf{Z}_2) \leq \sup|\Psi'(\cdot)| \times \sigma_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\sigma_{\max}(\mathbf{Z}_2)$.

(which we denote by $\|\cdot\|$) and is also known as the operator norm or the spectral norm since $\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}(A)$ is the largest singular value for A. For Lemma 1, we define an Asymptotic Contraction Mapping (ACM) as (see DS, p.841):

Definition 1 (ACM). *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. For $\omega \in \Omega$, let $W_1(\omega, \beta_0), \ldots, W_n(\omega, \beta_0)$ denote a sample of size n from $\mathbb{P}$, where $\beta_0 \in \mathbb{R}^k$ denotes a fixed parameter of interest. For each $n \geq 1$ and $\omega \in \Omega$, let $K^\omega(\cdot)$ be a function defined on a set $\mathcal{X}$ where $(\mathcal{X}, d)$ is a metric space. The collection $\{K_n^\omega(\cdot) : n \geq 1, \omega \in \Omega\}$ is an asymptotic contraction mapping (ACM) on $(\mathcal{X}, d)$ if there exists a constant $c \in [0, 1)$ that does not depend on n or ω, and sets $\{A_n\}$ with $A_n \subset \Omega$ and $\mathbb{P}(A_n) \to 1$ as $n \to \infty$, such that for each $\omega \in A_n$, $K_n^\omega(\cdot)$ maps $\mathcal{X}$ to itself and for all $x, y \in \mathcal{X}$, $d(K_n^\omega(x), K_n^\omega(y)) \leq c \times d(x, y)$.*

We can now establish that:

Lemma 1. *If Assumptions 1 and 2(i) hold, there exists and open ball centered at β_0 with closure $B_0 \subset \mathcal{N}$ such that $\{M_n(\phi)\}$ is an ACM on $(B_0, \|\cdot\|)$.*

Proof. Take ϕ and $\gamma \in B_0 \subset \mathcal{N}$. By the Mean Value Theorem we have:

$$M_n(\phi) = M_n(\gamma) + \nabla_\phi M_n(\phi^*)(\phi - \gamma),$$

where ϕ^* is (componentwise) between ϕ and γ. Notice that

$$\nabla_\phi M_n(\phi) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \nabla_\phi \mathbf{Y}(\phi),$$

where $\nabla_\phi \mathbf{Y}(\phi)_{n \times k} = [\nabla_\phi y_1(\phi) \ldots \nabla_\phi y_n(\phi)]^\top$ with $\nabla_\phi y_i(\phi) \equiv \Psi_0'(z_{1i} - \phi^\top z_{2i}) z_{2i}$. Then, adding and subtracting $(\nabla_\phi M_n(\beta_0) - \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i \nabla y_i(\beta_0)^\top] + \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i \nabla y_i(\phi^*)^\top])$ $\times(\phi - \gamma)$, we can obtain:

$$\begin{aligned} &\|M_n(\phi) - M_n(\gamma)\| = \|\nabla_\phi M_n(\phi^*)(\phi - \gamma)\| \\ &\leq \|\nabla_\phi M_n(\beta_0)(\phi - \gamma)\| + \|\mathbb{E}[x_i x_i^\top]^{-1} \left\{\mathbb{E}[x_i \nabla y_i(\phi^*)^\top] - \mathbb{E}[x_i \nabla y_i(\beta_0)^\top]\right\} (\phi - \gamma)\| + o_p(\|\phi - \gamma\|), \end{aligned} \tag{20}$$

where $\|\cdot\|$ is the Euclidean norm.

The o_p term in (20) corresponds to $\|\{-\left(\nabla_\phi M_n(\beta_0) - \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i \nabla y_i(\beta_0)^\top]\right) + \nabla_\phi M_n(\phi^*) - \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i \nabla y_i(\phi^*)^\top]\}(\phi - \gamma)\|$. This can be shown to converge to zero in probability if $\nabla_\phi M_n(\cdot) - \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i \nabla y_i(\cdot)^\top]$ converges uniformly to zero in probability. This is established using Lemma 2.9 in Newey and McFadden 1994.

To employ this lemma, first note that the parameter space is compact by Assumption 1. Also by Assumption 1, $\|\nabla_\phi M_n(\phi) - \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i \nabla y_i(\phi)^\top]\|_2$ converges to zero in

probability for given ϕ by the Law of Large Numbers and the Continuous Mapping Theorem. Moreover, we can also obtain that

$$\|\nabla_\phi M_n(\phi) - \nabla_\phi M_n(\gamma)]\|_2 \leq \hat{B}_n \|\phi - \gamma\|,$$

where $\hat{B}_n = O_p(1)$. For notational simplicity, assume that $k = 1$. In this case, by the Mean Value Theorem,

$$\begin{aligned}
\nabla_\phi M_n(\phi) - \nabla_\phi M_n(\gamma) &= \left(\sum_{i=1}^n x_i^2\right)^{-1} \times \sum_{i=1}^n x_i z_{i2} \left(\Psi_0'(z_{1i} - \phi z_{2i}) - \Psi_0'(z_{1i} - \gamma z_{2i})\right) \\
&= -\left(\sum_{i=1}^n x_i^2\right)^{-1} \times \sum_{i=1}^n x_i z_{i2}^2 \Psi_0''(z_{1i} - \phi^* z_{2i})(\phi - \gamma),
\end{aligned}$$

where ϕ^* is (componentwise) between ϕ and γ. By the triangle inequality and Assumption 1 (v) (i.e., $|\Psi_0''(\cdot)| \leq c$ for some c), we have that

$$|\nabla_\phi M_n(\phi) - \nabla_\phi M_n(\gamma)| \leq \underbrace{c \left(\sum_{i=1}^n x_i^2/n\right)^{-1} \times \sum_{i=1}^n \left|x_i z_{i2}^2\right|/n}_{\equiv \hat{B}_n} \times |\phi - \gamma|.$$

Given Assumption 1 ii, $\hat{B}_n = O_p(1)$. Finally, continuity of $\Psi_0'(\cdot)$ implies continuity of $\mathbb{E}[x_i x_i^\top]^{-1} \times \mathbb{E}[x_i \nabla y_i(\phi)^\top]$ for $\phi \in \mathcal{N}$. These suffice for uniform convergence by Lemma 2.9 in Newey and McFadden 1994.

Since $\Psi_0(\cdot) \in C_2$, $\mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i \nabla y_i(\cdot)^\top]$ is continuously differentiable, which allows one to control the second term in (20) by the diameter of B_0. To see this, and again using $k = 1$ for notational simplicity, note that

$$\begin{aligned}
\left|\mathbb{E}[x_i \nabla y_i(\phi^*)^\top] - \mathbb{E}[x_i \nabla y_i(\beta_0)^\top]\right| &= \left|\mathbb{E}[x_i z_{i2}^2 \Psi_0''(z_{1i} - \tilde{\phi} z_{2i})]\right| \times |(\phi^* - \beta_0)| \\
&\leq c \times \mathbb{E}[|x_i z_{i2}^2|] \times |(\phi^* - \beta_0)|.
\end{aligned}$$

where $\tilde{\phi}$ is a value between ϕ^* and β_0 and the second line uses the triangle inequality and Assumption 1 (v) (i.e., $|\Psi_0''(\cdot)| \leq c$ for some c). The second term in (20) can thus be seen to be $o(\sup_{\phi,\gamma \in B_0} |\phi - \gamma|)$.

Then, as noted in the proof for Lemma 5 in DS, it suffices to show that $\mathbb{P}(\|\nabla_\phi M_n(\beta_0)(\phi - \gamma)\| < C\|(\phi - \gamma)\|) \to 1$ as $n \to \infty$, where C is a positive scalar such that $C < 1$. Since

the matrix norm induced by the Euclidean distance is submultiplicative, we have that:

$$\|\nabla_\phi M_n(\beta_0)(\phi-\gamma)\| \leq \|\nabla_\phi M_n(\beta_0)\|_2 \times \|(\phi-\gamma)\| = \sigma_{\max}(\nabla_\phi M_n(\beta_0)) \times \|(\phi-\gamma)\|.$$

Since $\mathbb{P}(\sigma_{\max}(\nabla_\phi M_n(\beta_0)) < 1) \to 1$ as $n \to \infty$, we obtain the desired result. □

For the uniform convergence condition, we obtain that:

Lemma 2. *If Assumptions 1 and 2(ii) hold,* $\sup_\phi \|\widehat{M}_n(\phi) - M_n(\phi)\| = o_p(1)$ *as* $n \to \infty$.

Proof. First notice that

$$\begin{aligned} y_i(\phi) &= y_i - \Psi_0(z_{1i} - \phi^\top z_{2i}) - \epsilon_i \\ &= (x_i^\top \beta_0 + \Psi_0(z_{1i} - z_{2i}^\top \beta_0) + \epsilon_i) - \Psi_0(z_{1i} - \phi^\top z_{2i}) - \epsilon_i \\ &= x_i^\top \beta_0 + \Big(\Psi_0(z_{1i} - z_{2i}^\top \beta_0) - \Psi_0(z_{1i} - \phi^\top z_{2i})\Big) \end{aligned}$$

and

$$\begin{aligned} \hat{y}_i(\phi) &= (x_i^\top \beta_0 + \Psi_0(z_{1i} - z_{2i}^\top \beta_0) + \epsilon_i) - \hat{\Psi}(z_{1i} - \phi^\top z_{2i}) \\ &= x_i^\top \beta_0 + (\Psi_0(z_{1i} - z_{2i}^\top \beta_0) - \Psi_0(z_{1i} - \phi^\top z_{2i})) + \epsilon_i + (\Psi_0(z_{1i} - \phi^\top z_{2i}) - \hat{\Psi}(z_{1i} - \phi^\top z_{2i})) \\ &= y_i(\phi) + \epsilon_i + (\Psi_0(z_{1i} - \phi^\top z_{2i}) - \hat{\Psi}(z_{1i} - \phi^\top z_{2i})). \end{aligned}$$

Consequently, we can write:

$$\widehat{\mathbf{Y}}(\phi)_{n\times 1} = \mathbf{Y}(\phi)_{n\times 1} + \epsilon_{n\times 1} + \tilde{\Psi}(\phi, \mathbf{Z})_{n\times 1},$$

where $\epsilon = [\epsilon_1 \ \cdots \ \epsilon_n]^\top$ and $\tilde{\Psi}(\phi, \mathbf{Z}) = [(\Psi_0(z_{11} - \phi^\top z_{21}) - \hat{\Psi}(z_{11} - \phi^\top z_{21})) \ \cdots \ (\Psi_0(z_{1n} - \phi^\top z_{2n}) - \hat{\Psi}(z_{1n} - \phi^\top z_{2n}))]^\top$. ($\widehat{\mathbf{Y}}(\phi)$ and $\mathbf{Y}(\phi)$ were defined earlier.) Using the definitions for $M_n(\phi)$ and $\widehat{M}_n(\phi)$ in equations (18) and (19), we then obtain:

$$\widehat{M}_n(\phi) - M_n(\phi) = \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon}_{\equiv A} + \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\Psi}(\phi, \mathbf{Z})}_{\equiv B(\phi)}.$$

The first term above (A) does not depend on ϕ and converges in probability to 0 given Assumption 1. (This obtains from $A = (\sum_{i=1}^n x_i x_i^\top/n)^{-1} \times (\sum_{i=1}^n x_i \epsilon_i^\top/n) \to_p \mathbb{E}[x_i x_i^\top]^{-1}\mathbb{E}[x_i \epsilon_i] = 0$.)

For the second term, let $B_1 \equiv \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\sqrt{n}$ and $B_2(\phi) \equiv \tilde{\Psi}(\phi, \mathbf{Z})/\sqrt{n}$ so that

$B(\phi)_{k\times 1} = B_1^\top B_2(\phi)$. Then, it can be deduced that

$$\|B(\phi)\|^2 = \|B_1^\top B_2(\phi)\|^2 = (B_1(:,1)^\top B_2(\phi)) + \cdots + (B_1(:,k)^\top B_2(\phi)),$$

where $B_1(:,j)$ is the jth column of B_1. Applying the Cauchy-Schwatz inequality to each of the k terms in the sum, we get:

$$\|B(\phi)\|^2 \leq \|B_1(:,1)\|^2\|B_2(\phi)\|^2 + \cdots + \|B_1(:,k)\|^2\|B_2(\phi)\|^2 = (\|B_1(:,1)\|^2 + \cdots + \|B_1(:,k)\|^2)\|B_2(\phi)\|^2.$$

Notice that $\|B_1(:,1)\|^2 + \cdots + \|B_1(:,k)\|^2 = \|B_1\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm. Since $\|B_1\|_F^2 = tr(B_1^\top B_1)$, where $tr(\cdot)$ is the trace operator, we have that

$$\|B_1\|_F^2 = tr((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}n) = tr((\mathbf{X}^\top\mathbf{X})^{-1}n) = tr((\sum_{i=1}^{n} x_i x_i^\top/n)^{-1}).$$

Given that $\sum_{i=1}^n x_i x_i^\top/n \to_p \mathbb{E}[x_i x_i^\top]$, continuity of $tr(\cdot)$ and non-singularity of $\mathbb{E}[x_i x_i^\top]$, $\|B_1\|_F^2 \to_p tr(\mathbb{E}[x_i x_i^\top]^{-1})$.

In addition, notice that:

$$\|B_2(\phi)\|^2 = B_2(\phi)^\top B_2(\phi) = \tilde{\Psi}(\phi,\mathbf{Z})^\top\tilde{\Psi}(\phi,\mathbf{Z})/n = \sum_{i=1}^{n}\left(\Psi_0(z_{1i} - \phi^\top z_{2i}) - \hat{\Psi}(z_{1i} - \phi^\top z_{2i})\right)^2/n.$$

Since $\left(\Psi_0(z_{1i} - \phi^\top z_{2i}) - \hat{\Psi}(z_{1i} - \phi^\top z_{2i})\right)^2 \leq \sup_\phi |\Psi_0(z_{1i} - \phi^\top z_{2i}) - \hat{\Psi}(z_{1i} - \phi^\top z_{2i})|^2 = \sup|\Psi_0(\cdot) - \hat{\Psi}(\cdot)|^2$, $\|B_2(\phi)\|^2 \leq \sup|\Psi_0(\cdot) - \hat{\Psi}(\cdot)|^2$ which converges to zero in probability if $\hat{\Psi}(\cdot)$ converges to $\Psi_0(\cdot)$ uniformly in probability. Thus, $\|B_2(\phi)\|^2 \to_p 0$ and consequently

$$\|B(\phi)\|^2 \leq \|B_1\|_F^2 \times \|B_2(\phi)\|^2 \to_p 0$$

uniformly in ϕ. □

C Implementing the NPR estimator

As explained in subsection 2.2, the NPR estimator consists of the following steps:

1. Pick initial parameter values $(\hat{\beta}_{k0}, \hat{\beta}_{l0})$.
2. For iteration j, calculate $Z_{it} = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \hat{\beta}_{kj-1}k_{it+1} - \hat{\beta}_{lj-1}\mathbb{E}_{it}[l_{it+1}|\Omega_{it}]$.
3. Fit the model $y_{it} = \beta_k k_{it} + \beta_l l_{it} + \Psi(Z_{it}) + \epsilon_{it}$ using the shape constrained estimation protocol of Pya and Wood (2015) to obtain $(\hat{\beta}_j, \hat{\Psi}_j)$.

4. Calculate the Euclidean distance between $(\hat{\beta}_{kj}, \hat{\beta}_{lj})$ and $(\hat{\beta}_{kj-1}, \hat{\beta}_{lj-1})$. If the distance is below some tolerance level, stop and treat $\hat{\beta}_j$ as the model's parameter estimates. If not then update the iteration number $j \leftarrow j+1$ and repeat from step 2.

For more general production functions, one instead should use $Z_{it} = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \int f(k_{it+1}, l_{it+1}; \hat{\beta}) dF_{it}(l_{it+1}))$ in step 2 and $y_{it} = f(k_{it}, l_{it}; \beta) + \Psi(Z_{it}) + \epsilon_{it}$ in step 3.

Implementing this iterative algorithm requires a number of decisions. First, one must decide on the initialization values $(\hat{\beta}_{k0}, \hat{\beta}_{l0})$. Ichimura and Todd (2007) recommend that, when fitting a generalized additive model, initial values be obtained from a linear regression of the outcome variable on all dependent variables (i.e. both those that enter the model linearly and those that are arguments of non-parametric smooth model components). In our context this would involve regressing y_{it} on a constant, k_{it}, l_{it} and $Z_{it} \equiv \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \beta_{k0} k_{it+1} - \beta_{l0} \mathbb{E}[l_{it+1}|\Omega_{it}]$, which nonetheless depends on the unknown parameters (β_{k0} and β_{l0}). One alternative is to instead regress y_{it} on a constant and $X_{it} = [k_{it}, l_{it}, \mathbb{E}_{it}[y_{it+1}|\Omega_{it}], k_{it+1}, \mathbb{E}[l_{it+1}|\Omega_{it}]]$ and set $(\hat{\beta}_{k0}, \hat{\beta}_{l0})$ as the coefficient estimates on current-period capital and labor respectively. However, in both Monte Carlo simulations and our empirical context, this approach frequently initialized the capital coefficient at a negative value due to high correlations between y_{it} and $\mathbb{E}_{it}[y_{it+1}]$ and between k_{it} and k_{it+1}, which led the NPR estimator to converge on nonsensical parameter values (e.g. $\hat{\beta}_k < 0$).

Rather than specify a rule for selecting a single initialization parameter vector we therefore recommend a grid search approach. Under this method, one specifies a grid of N initial values for $(\hat{\beta}_{k0}, \hat{\beta}_{l0})$ and obtains N parameter estimates by implementing the NPR estimator initialised at each point on the grid. The optimal parameter estimates are then chosen as those associated with the lowest least squares objective function from the shape constrained estimation protocol. All results in this paper were calculated using this initialization approach implemented on a 16-point initialization grid consisting of all combinations of $\hat{\beta}_{k0} \in [0.05, 0.333, 0.617, 0.9]$ and $\hat{\beta}_{l0} \in [0.05, 0.333, 0.617, 0.9]$.[39]

The selection of initial values of production function input coefficients is also required by the ACF estimator, whose Monte Carlo results were created by initializing at the true values "to ease non-linear search". ACF note their results are "fairly robust to other nearby starting values (e.g OLS estimates), though with further away starting values, one can sometimes end up at the spurious minimum around $(\beta_k = 0, \beta_l = \beta_l + \beta_k)$", as described in

39. This does not constrain the algorithm to converge on an estimate within the [0.05,0.9] range. Unlike ACF whose estimates can be below zero and above 1, the Monte Carlo and empirical application for NPR do not end up outside economically sensible ranges.

their footnote 16. In our empirical context, we found that ACF estimates remained almost unchanged from their initial values when using OLS initialization combined with Nelder-Mead minimization, which is the default in the popular `'prodest'` package in `Stata` (Rovigatti and Mollisi 2018). Using the gradient-based BFGS minimization routine, however, achieved results that differed more from the OLS estimates and achieved a lower value of the objective function. The ACF results we report were therefore obtained using OLS initialization and BFGS minimization.

The protocol outlined by Pya and Wood (2015) (PW) allows one to estimate a general additive model of the form

$$y_i = \theta^\top x_i + \sum_j f_j(z_{ji}) + \epsilon_i, \tag{21}$$

where y is a univariate response variable, x is a vector of linear independent variables and θ a vector of unknown parameters. The non-linear part of the model is represented by unknown, monotonically increasing smooth predictor functions f_j, with predictor variable z_j. The variable ϵ is an unobserved mean-zero disturbance.[40] To estimate this model, PW follow convention in the literature on general additive model estimation and approximate the unknown smooth functions f_j with penalized B-splines (i.e. P-splines). Under this simplification, the model becomes

$$y_i = \theta^\top x_i + \sum_{j=1}^{q} \gamma_j B_j(z_i) + \epsilon_i, \tag{22}$$

where q is the number of basis functions, B_j are B-spline basis functions of at least second order that represent smooth functions over interval $[a, b]$ based on evenly-spaced knots and γ_j are the corresponding spline coefficients. Ensuring the smooth part of the model is monotonically increasing amounts to imposing restrictions on the spline parameters γ and PW show how the model in equation (22) can be reparameterized to guarantee such restrictions are satisfied. Following this reparameterization, estimates of the original model parameters can be obtained via minimising the difference between the response variable and the observable components on the RHS of equation (22). To avoid overfitting during this minimization, one should also include a penalty term that controls the 'wiggliness' of the B-splines. Specifically, PW account for this concern by penalizing the squared difference between adjacent B-spline parameters. While it is possible to pre-

40. While we focus this description of Pya and Wood (2015) on our particular application for parsimony, it should be noted that their framework allows for more general models.

specify the 'smoothing parameter', which scales 'wiggliness' term in the minimization and thereby controls the smoothness of the estimated functions, PW propose to estimate the 'optimal' smoothing parameter via an outer estimation algorithm. The outer part of the estimation uses the generalized cross validation prediction error criterion to evaluate the performance of the model estimated using a particular value of the smoothing parameter and finds the optimal smoothing parameter according to this criterion using the Newton-Raphson method.

In our application of the PW estimator, we specify the smooth term Ψ using monotone increasing P-splines (using the `'bs="mpi"'` option), with 20 basis points. The smoothing parameter is estimated using the Newton-Raphson method, while the model coefficients are estimated using the BFGS algorithm.

D Extensions to the NPR estimator

D.1 Value added production functions

Adapting the NPR estimator to a value added production function is trivial if one has data on firms' expected year-ahead value added. In this case, one can simply change y in the estimation algorithm from log turnover to log value added and the resultant estimates will have a value added interpretation. If such data is unavailable, but one instead possesses data on firms' expected material inputs in addition to expected turnover and employment, one can use copula methods to estimate expected log value added. Specifically, firms' subjective *joint* distribution between expected turnover and expected materials can be estimated by applying a parametric copula to firms' subjective marginal distributions.[41] One can then take an adequately large number of draws of expected turnover and intermediates from this joint subjective distribution, calculate $\ln(turnover - intermediates)$ for each draw and recover expected log value added as the mean over the random draws where this quantity is defined (i.e. when $(turnover - intermediates)$ is greater than zero). Equipped with an estimate of expected value added, one can again simply set y as log value added rather than log turnover and proceed with the iterative NPR estimation algorithm detailed in subsection 2.2. Although in theory possible, we do not explore this extension in our empirical application the data because expectations of material inputs were only elicited in the 2017 wave of

41. One could, for example, fit the normal copula to match the observed correlation between observed turnover and materials.

the MES, leading to small industry-specific samples.[42]

D.2 NPR Estimation with measurement error in expectations data ("EZ-NPR"

As discussed in subsection 2.4 of the main text, NPR is vulnerable in the presence of measurement error in firms' subjective expectations data. Inputs mismeasurement is well known to be a common problem for production function estimation, and can in principle cause substantial bias for proxy methods (e.g., Collard-Wexler and De Loecker 2021) and other approaches (e.g., Griliches and Mairesse 1999).

We proposed a generation called EZ-NPR which is based on moments given by equation (12), which we repeat here for reference

$$M_{it}^{x}(Z_{it}^{*}) = M_{it}^{x}(Z_{it}) - \frac{\partial^2 M_{it}^{x}(Z_{it})}{\partial \mathbb{E}_{it}[y_{it+1}]^2}\frac{\mathbb{E}[\upsilon_y^2]}{2} - \frac{\partial^2 M_{it}^{x}(Z_{it})}{\partial \mathbb{E}_{it}[l_{it+1}]^2}\frac{\mathbb{E}[\upsilon_l^2]}{2},$$

where

$$M_{it}^{x}(Z_{it}) = \left(y_{it} - \beta_k k_{it} - \beta_l l_{it} - \Psi\left(Z_{it}\right)\right) x_{it},$$

and

$$Z_{it} = \mathbb{E}_{it}[y_{it+1}|\Omega_{it}] - \beta_k k_{it+1} - \beta_l \mathbb{E}_{it}[l_{it+1}|\Omega_{it}].$$

Here we derive the corrected EZ-NPR moments and discuss how one can use them to recover parameter estimates using GMM.

D.2.1 Deriving the corrected moments

Constructing the corrected moments requires derivation and calculation of the second partial derivatives of M_{it}^{x} with respect to $\mathbb{E}_{it}[y_{it+1}]$ and $\mathbb{E}_{it}[l_{it+1}]$.

First, consider the case where the instrument x_{it} does not depend on either of the variables measured with error (e.g. $x_{it} \in \{1, l_{it}, k_{it}, k_{it+1}\}$). From equation (12) and the definition of Z_{it}, the product rule implies

$$\begin{aligned}\frac{\partial^2 M_{it}^{x}}{\partial \mathbb{E}_{it}[y_{it+1}]^2} &= -\Psi''(Z_{it})x_{it} \\ \frac{\partial^2 M_{it}^{x}}{\partial \mathbb{E}_{it}[l_{it+1}]^2} &= -\beta_l^2 \Psi''(Z_{it})x_{it},\end{aligned}$$

42. Copula methods can also be employed in (non-linear) gross output product functions with both labor and intermediates where the joint distribution for both variables is required.

where $\Psi''(Z_{it})$ denote the second derivative of the b-spline function with respect to the argument Z_{it}.

Now, consider the case where the instrument x_{it} *does* depend on the variables measured with error (e.g. $x_{it} = Z_{it}$). In this case we have:

$$\frac{\partial^2 M_{it}^x}{\partial \mathbb{E}_{it}[y_{it+1}]^2} = -\Psi''(Z_{it})x_{it} - 2\frac{\partial x_{it}}{\partial \mathbb{E}_{it}[y_{it+1}]}\Psi'(Z_{it}) + \frac{\partial^2 x_{it}}{\partial \mathbb{E}_{it}[y_{it+1}]^2}\left(y_{it} - \beta_k k_{it} - \beta_l l_{it} - \Psi(Z_{it})\right)$$
$$\frac{\partial^2 M_{it}^x}{\partial \mathbb{E}_{it}[l_{it+1}]^2} = -\beta_l^2\Psi''(Z_{it})x_{it} + 2\beta_l\frac{\partial x_{it}}{\partial \mathbb{E}_{it}[l_{it+1}]}\Psi'(Z_{it}) + \frac{\partial^2 x_{it}}{\partial \mathbb{E}_{it}[l_{it+1}]^2}\left(y_{it} - \beta_k k_{it} - \beta_l l_{it} - \Psi(Z_{it})\right),$$

where $\Psi''(Z_{it})$ denote the second derivative of the b-spline function with respect to the argument Z_{it}. For $x_{it} = Z_{it}$, $\partial x_{it}/\partial \mathbb{E}_{it}[y_{it+1}] = 1$, $\partial x_{it}/\partial \mathbb{E}_{it}[l_{it+1}] = -\beta_l$ and the second-order partial derivatives are both equal to zero.

The precise form of the correction terms evidently depends on how one models Ψ. Following the baseline NPR implementation, we model Ψ as a b-spline with monotonicity enforced via coefficient ordering constraints

$$\Psi(Z_{it}) = c_0 + \sum_{j=1}^{K} c_j\, b_j(Z_{it}), \qquad c_1 \le c_2 \le \cdots \le c_K,$$

where $b_1, \ldots, b_K$ are b-spline basis functions of degree p and c_0 is an unrestricted intercept.[43]

The first and second derivatives of Ψ with respect to Z_{it} are therefore

$$\Psi'(Z_{it}) = \sum_{j=1}^{K} c_j\, b_j'(Z_{it})$$
$$\Psi''(Z_{it}) = \sum_{j=1}^{K} c_j\, b_j''(Z_{it}),$$

which can be computed directly from the derivatives of the b-spline basis functions.

Putting all these components together, the corrected moments, which we denote $\tilde{M}_{it}^x$

43. The ordering constraint $c_{j+1} \ge c_j$ is sufficient for monotonicity because the derivative of a degree-p b-spline can be written as

$$\Psi'(Z_{it}) = p\sum_j \frac{c_{j+1} - c_j}{\tau_{j+p} - \tau_j}\tilde{b}_j(Z_{it}),$$

where $\tilde{b}_j$ are b-splines of degree $p-1$ (which are everywhere non-negative) and τ_j are the knot positions. Since $c_{j+1} - c_j \ge 0$ and $\tau_{j+p} - \tau_j > 0$, we have $\Psi'(Z_{it}) \ge 0$ everywhere.

are therefore

$$\tilde{M}_{it}^{x} = \left(y_{it} - \beta_k k_{it} - \beta_l l_{it} - c_0 - \sum_{j=1}^{K} c_j \, b_j(Z_{it}) \right) x_{it} + x_{it} \left(\frac{\mathbb{E}[\upsilon_y^2]}{2} + \beta_l^2 \frac{\mathbb{E}[\upsilon_l^2]}{2} \right) \sum_{j=1}^{K} c_j \, b_j''(Z_{it}), \tag{23}$$

for $x \in \{1, k_{it}, l_{it}, k_{it+1}\}$ and

$$\begin{aligned} \tilde{M}_{it}^{x} = & \left(y_{it} - \beta_k k_{it} - \beta_l l_{it} - c_0 - \sum_{j=1}^{K} c_j \, b_j(Z_{it}) \right) x_{it} \\ & + \left(\frac{\mathbb{E}[\upsilon_y^2]}{2} + \beta_l^2 \frac{\mathbb{E}[\upsilon_l^2]}{2} \right) \left(x_{it} \sum_{j=1}^{K} c_j \, b_j''(Z_{it}) + 2 \sum_{j=1}^{K} c_j \, b_j'(Z_{it}) \right), \end{aligned} \tag{24}$$

for $x = Z_{it}$.

D.2.2 Implementing EZ-NPR

The EZ-NPR estimator is essentially GMM estimation using moments $\mathbb{E}[\tilde{M}_{it}^{x}] = 0$, where $\tilde{M}_{it}^{x}$ is given by equations (23) and (24). A first implementation decision regards the choice of instruments. In principle, one could derive corrected moments similar to equation (24) with $\mathbb{E}_{it}[y_{it+1}]$ and $\mathbb{E}_{it}[l_{it+1}]$ as instruments. In practice, however, we found that including Z_{it} as the only instrument contaminated by measurement error performed well in DGPs that featured biased expectations, as well as those that featured measurement error. This is because the Z_{it} instrument corrects moments by a single variance parameter (the $\left(\mathbb{E}[\upsilon_y^2]/2 + \beta_l^2 \mathbb{E}[\upsilon_l^2]/2\right)$ in equation (24)), which can equally absorb the variance of expectational biases when such biases are present. When either $\mathbb{E}_{it}[y_{it+1}]$ or $\mathbb{E}_{it}[l_{it+1}]$ is included in the instrument set, the corrected moments separately identify $\mathbb{E}[\upsilon_y^2]/2$ and $\mathbb{E}[\upsilon_l^2]/2$. While this improves performance in DGPs with measurement error, the true parameters are no longer recovered in DGPs with biased expectations as the resultant estimator attempts to decompose the single bias variance into distinct components that load separately on $\mathbb{E}_{it}[y_{it+1}]$ and $\mathbb{E}_{it}[l_{it+1}]$. A further practical advantage of Z_{it} is that, as a known function of β_l and β_k, it sharpens the GMM objective in these directions. We therefore take $\{1, k_{it}, l_{it}, k_{it+1}, Z_{it}\}$ as the instrument set we use to implement the EZ-NPR estimator.

The EZ-NPR moments feature $K+4$ parameters: the two production function factor elasticities (β_l, β_k), the intercept c_0 and K spline coefficients $(c_1, \ldots, c_K)$ of the Ψ func-

tion, and the combined measurement error variance $\gamma \equiv \frac{\mathbb{E}[v_y^2]}{2} + \beta_l^2 \frac{\mathbb{E}[v_l^2]}{2}$. In Monte Carlo experiments, we found that numerical optimisers failed to recover the true parameter estimates when optimising over all $K+4$ parameters simultaneously. This is unsurprising: the flexibility of the b-spline function introduces local optima in the spline coefficients, and these interact with the measurement error parameter, whose identification relies on the shape of Ψ. We therefore reduce the dimension of the outer optimisation to two by concentrating out c_0, $(c_1, \ldots, c_K)$, and γ at each candidate (β_l, β_k).

We now describe the full estimation algorithm, which proceeds in three stages: (i) knot selection, (ii) corrected GMM with identity weighting and (iii) efficient two-step GMM.

Stage 1: Knot selection. The b-spline basis functions require a set of interior knots $\kappa_1 < \cdots < \kappa_J$, which determine where the basis functions are centred. Since the argument Z_{it} depends on (β_l, β_k), the distribution of Z_{it} — and hence the appropriate knot locations — changes with the candidate parameters. Recomputing knots at each evaluation of the objective would introduce discontinuities, which appeared to compromise consistency in Monte Carlo simulations. We therefore fix the knots once, prior to the corrected estimation, using the following procedure. For each point on a (β_l, β_k) grid,[44] we evaluate the uncorrected GMM objective (i.e. setting $\gamma = 0$) using the identity weighting matrix. The grid point achieving the lowest objective is polished using Nelder-Mead optimisation and the resulting estimates $(\hat{\beta}_l^{(0)}, \hat{\beta}_k^{(0)})$ are used to construct $\hat{Z}_{it}^{(0)} = \mathbb{E}_{it}[y_{it+1}] - \hat{\beta}_k^{(0)} k_{it+1} - \hat{\beta}_l^{(0)} \mathbb{E}_{it}[l_{it+1}]$. The interior knots are set at equally spaced quantiles of $\hat{Z}_{it}^{(0)}$ and remain fixed for all subsequent stages of the estimation.

Stage 2: Corrected GMM. For a given candidate (β_l, β_k), the intercept c_0 and spline coefficients $(c_1, \ldots, c_K)$ can be recovered by regressing $r_{it} \equiv y_{it} - \beta_k k_{it} - \beta_l l_{it}$ on the b-spline basis functions $b_1(Z_{it}), \ldots, b_K(Z_{it})$ and a constant, subject to the monotonicity constraints $c_1 \leq c_2 \leq \cdots \leq c_K$. When Z_{it} is subject to classical measurement error, however, the resultant c estimates will be attenuated, since the b-spline basis functions are evaluated at the noisy Z_{it} rather than the true Z_{it}^*. The regression must therefore be adjusted using the second-order Taylor expansion principle of EZ, which corrects the normal equations underlying the constrained regression to account for measurement error in the argument of the basis functions. This creates a circularity: if the measurement

44. Similar to the baseline NPR estimator, the range of this grid is 0.1 to 0.9. Unlike the baseline NPR estimator, however, MC simulations indicated the importance of having a sufficiently fine grid. We therefore use a 15×15 grid, which consists of 15 equally-spaced points between 0.1 and 0.9.

error variance γ were known, the correction would have a closed-form solution, but to concentrate out the measurement error variance one requires unbiased estimates of c_0 and $(c_1, \ldots, c_K)$. For given (β_l, β_k), we therefore concentrate out c_0, $(c_1, \ldots, c_K)$ and γ using the following iterative approach:

1. Initialise the measurement error variance at $\gamma^{(0)} = 0$ and set the iteration counter $\iota = 0$.

2. Given $\gamma^{(\iota)}$, solve the corrected constrained regression for c_0 and $(c_1, \ldots, c_K)$. This involves solving the quadratic programme

$$\min_{c_0, c_1, \ldots, c_K} \frac{1}{2} \boldsymbol{c}' A^{(\iota)} \boldsymbol{c} - \boldsymbol{f}^{(\iota)\prime} \boldsymbol{c} \qquad \text{s.t.} \quad c_1 \leq c_2 \leq \cdots \leq c_K,$$

   where $\boldsymbol{c} = (c_0, c_1, \ldots, c_K)'$ and the corrected normal equation components are

$$A^{(\iota)} = \frac{B'B}{n} + \gamma^{(\iota)} \frac{B_2'B + 2B_1'B_1 + B'B_2}{n}, \qquad \boldsymbol{f}^{(\iota)} = \frac{B'r}{n} + \gamma^{(\iota)} \frac{B_2'r}{n}.$$

   Here B denotes the matrices of b-spline basis function values evaluated at the observed Z_{it} augmented with a column of ones for the intercept, and B_1 and B_2 denote the corresponding matrices of first and second derivatives (with a leading column of zeros for the intercept).[45] When $\gamma^{(0)} = 0$, this reduces to an uncorrected constrained regression.

3. Using the solution from step 2, compute the first and second derivatives of the estimated Ψ function:

$$\hat{\Psi}'(Z_{it}) = \sum_{j=1}^{K} \hat{c}_j^{(\iota)} \, b_j'(Z_{it}), \qquad \hat{\Psi}''(Z_{it}) = \sum_{j=1}^{K} \hat{c}_j^{(\iota)} \, b_j''(Z_{it}).$$

4. Compute the residuals $\hat{\epsilon}_{it} = r_{it} - \hat{c}_0^{(\iota)} - \sum_{j=1}^{K} \hat{c}_j^{(\iota)} b_j(Z_{it})$ and the observed (uncorrected) moment vector $\bar{\boldsymbol{g}}_{\text{obs}} = n^{-1} \sum_i \hat{\epsilon}_{it} \cdot \boldsymbol{h}_{it}$, where $\boldsymbol{h}_{it}$ is the vector of instruments. Using $\hat{\Psi}'$ and $\hat{\Psi}''$ from step 3, construct the correction vector $\boldsymbol{X}_{\text{corr}}$, whose m-th element is the sample mean of the second-derivative correction term from equations (23)–(24) corresponding to instrument x_{it}^m.

45. The correction terms arise from applying the EZ second-order Taylor expansion to each normal equation $\mathbb{E}[(r - \Psi(Z))\, b_k(Z)] = 0$. The product rule applied to $\Psi(Z) \cdot b_k(Z)$ generates three terms — $B_2'B$ from the curvature of Ψ times the level of the instrument, $2B_1'B_1$ from the slope of Ψ times the slope of the instrument, and $B'B_2$ from the level of Ψ times the curvature of the instrument.

5. Profile γ by non-negative weighted least squares of the observed moment vector on the correction vector:

$$\gamma_{\text{raw}}^{(\iota+1)} = \max\left(\frac{\boldsymbol{X}'_{\text{corr}} W \, \bar{\boldsymbol{g}}_{\text{obs}}}{\boldsymbol{X}'_{\text{corr}} W \, \boldsymbol{X}_{\text{corr}}},\ 0 \right),$$

where W is the GMM weighting matrix and the non-negativity constraint reflects the fact that γ is a variance.

6. Update with dampening: $\gamma^{(\iota+1)} = (1-\lambda)\gamma^{(\iota)} + \lambda\, \gamma_{\text{raw}}^{(\iota+1)}$, where $\lambda \in (0,1)$ is a dampening factor. Increment ι and return to step 2. Repeat until $|\gamma^{(\iota+1)} - \gamma^{(\iota)}| < \tau$ for some tolerance τ, or until a maximum number of iterations is reached.

Upon convergence, the corrected moments can be evaluated at the final estimates of $\hat{c}_0$, $(\hat{c}_1, \ldots, \hat{c}_K)$ and $\hat{\gamma}$ using equations (23)–(24), and the GMM objective $Q(\beta_l, \beta_k) = n\, \bar{\boldsymbol{g}}'_{\text{corr}} W \, \bar{\boldsymbol{g}}_{\text{corr}}$ evaluated, where $\bar{\boldsymbol{g}}_{\text{corr}} = \bar{\boldsymbol{g}}_{\text{obs}} - \hat{\gamma}\, \boldsymbol{X}_{\text{corr}}$.

We implement this procedure using the identity weighting matrix $W = I$ on each point of the 4×4 grid, evaluate the GMM objective function and then "polish" the initialisation point that achieved the lowest objective value using Nelder-Mead optimisation.

Stage 3: Efficient two-step GMM. Using the estimates recovered in Stage 2, the efficient weighting matrix is then constructed as $W^* = \hat{S}^{-1}$, where $\hat{S} = n^{-1} \sum_i \hat{\boldsymbol{g}}_{it} \hat{\boldsymbol{g}}'_{it}$ is estimated from the first-step residuals. This efficient weighting matrix places greater weight on the precisely estimated predetermined moments relative to the noisier Z_{it} moment, thereby improving the identification of (β_l, β_k). The full procedure described in Stage 2 in then repeated using W^* rather than $W = I$ and the estimates it returns are taken as the EZ-NPR estimates.

D.3 The NPR estimator with biased expectations

In this section we consider how biases in firms' expectations affect the NPR estimator. We show that a particular type of bias in expected inputs and outputs provides an additional moment that can be used for identification and present extensions to the baseline NPR estimation algorithm that can accommodate biases in firms' productivity expectations. We also show how similar extensions can accommodate imperfect knowledge of the production technology among firms.

Case 1: Expected Input Bias and Technology-Consistent Output Bias

As stated in Section 2.5, the baseline NPR estimator can accommodate biased output expectations as long as such bias is also reflected in firms' expected inputs. Specifically, in the case of a firm with over-optimistic output expectations, for example, we require bias in the firm's employment expectations such that the integral of the production function with respect to expected labor equals the biased output expectation. In this particular case, and under the assumption of Cobb-Douglas technology, the difference between expected log output and realised log output gives:

$$\begin{aligned}
\mathbb{E}_{it-1}\left[y_{it}|\Omega_{it-1}\right] - y_{it} &= \beta_0 + \beta_k k_{it} + \beta_l \mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] + \mathbb{E}_{it-1}\left[\omega_{it}|\Omega_{it-1}\right] \\
&\quad - \beta_0 - \beta_k k_{it} - \beta_l l_{it} - g(\omega_{it-1}) - \xi_{it} - \epsilon_{it} \\
&= \beta_l\left(\mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] - l_{it}\right) + \mathbb{E}_{it-1}\left[\omega_{it}|\Omega_{it-1}\right] - g(\omega_{it-1}) - \xi_{it} - \epsilon_{it} \\
&= \beta_l\left(\mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] - l_{it}\right) + g(\omega_{it-1}) - g(\omega_{it-1}) - \xi_{it} - \epsilon_{it} \\
&= \beta_l\left(\mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] - l_{it}\right) - \xi_{it} - \epsilon_{it}.
\end{aligned} \tag{25}$$

Taking expectations of equation (25) and maintaining the assumption that firms' productivity expectations are unbiased, the errors (ξ, ϵ) cancel to zero:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{it-1}\left[y_{it}|\Omega_{it-1}\right] - y_{it}\right] &= \mathbb{E}\left[\beta_l\left(\mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] - l_{it}\right) - \xi_{it} - \epsilon_{it}\right] \\
&= \beta_l \mathbb{E}\left[\mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] - l_{it}\right] - \mathbb{E}\left[\xi_{it} - \epsilon_{it}\right] \\
&= \beta_l \mathbb{E}\left[\mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] - l_{it}\right].
\end{aligned}$$

Assuming $\mathbb{E}\left[\mathbb{E}_{it-1}\left[x_{it}|\Omega_{it-1}\right] - x_{it}\right] \neq 0$ for $x \in (y, l)$, the labor coefficient is identified as the average of sample mean expectation errors:

$$\beta_l = \frac{\mathbb{E}\left[\mathbb{E}_{it-1}\left[y_{it}|\Omega_{it-1}\right] - y_{it}\right]}{\mathbb{E}\left[\mathbb{E}_{it-1}\left[l_{it}|\Omega_{it-1}\right] - l_{it}\right]}. \tag{26}$$

This shows that, in the Cobb-Douglas technology case, a simple Wald estimator for β_l may therefore be obtained by replacing the sample mean expectation errors on the right hand side of equation (26) with their sample equivalents. Since the capital coefficient remains unidentified, equation (26) is not, on its own, enough to identify the entire production function. Rather it can be used either as a method to test the labor coefficient estimated by the baseline NPR algorithm or as an additional moment

in estimation. Nonetheless, it is worth reiterating that equation (26) is only informative if firms' expectations over inputs and output exhibit bias that is consistent with the production technology and that their expectations over productivity are unbiased. In our application using MES data, the expectations of sales and employment seem reasonably well aligned with outturns (see Table 7).

Case 2: Time-invariant Expected Productivity Bias

Suppose firms' productivity bias was time-invariant, so that $\iota_{it} = \iota_i$ in (13). If one had panel observations of firms' forecast errors, it would be possible to estimate firms' time-invariant bias and then use $\hat{\iota}_i$ to recover consistent estimates of β. In practice, this would lead to an 'outer' estimation loop and the NPR algorithm would become

1. Pick an initial vector of bias terms (one for each firm), $\hat{\iota}_0$.
2. For iteration r, implement the NPR estimation algorithm using $\left(\mathbb{E}_{it}[y_{it+1}|I_{it}] - \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) - \hat{\iota}_{r-1}\right)$ as the argument of the smooth function Ψ, to obtain coefficient estimates $\hat{\beta}_r$.
3. Use $\hat{\beta}_r$ to calculate $\mathrm{a}_{it} = \left(\mathbb{E}_{it-1}[y_{it}|I_{it-1}] - \int f(k_{it}, l_{it}; \hat{\beta}_r) dF_{it-1}(l_{it})\right)$ and $\mathrm{b}_{it} = y_{it} - f(k_{it}, l_{it}; \hat{\beta}_r)$ and their difference for all periods for which both expectations (needed to calculate a), and outcomes (needed to calculate b), are observed.
4. Recover an updated estimate of the bias terms $\hat{\iota}_r$ as the firm-level mean of $(\mathrm{a}_{it} - \mathrm{b}_{it})$ and compare to $\hat{\iota}_{r-1}$. If the difference is sufficiently small the stop, if not then repeat from step 2.

To understand how the calculation in step 3 is used to recover an estimate of ι_i, observe that

1. $\mathrm{a}_{it} = \mathbb{E}_{it-1}[y_{it}|I_{it-1}] - \int f(k_{it}, l_{it}; \beta) dF_{it-1}(l_{it}) = g(\omega_{it-1}) + \iota_i$
2. $\mathrm{b}_{it} = y_{it} - f(k_{it}, l_{it}; \beta) = g(\omega_{it-1}) + \xi_{it} + \epsilon_{it}$

therefore

$$\begin{aligned}(\mathrm{a}_{it} - \mathrm{b}_{it}) &= (g(\omega_{it-1}) + \iota_i) - (g(\omega_{it-1}) + \xi_{it} + \epsilon_{it}) \\ &= \iota_i - \xi_{it} - \epsilon_{it}\end{aligned}$$

Since the shocks ξ and ϵ are each assumed i.i.d. and mean-zero, multiple observations of productivity forecast errors $(\mathrm{a}_{it} - \mathrm{b}_{it})$ contain information on the time-invariant bias term ι.

Case 3: Time-varying Expected Productivity Bias

If firms' productivity bias is assumed to vary over time, one can proceed by specifying the bias as a function of observable firm attributes. Suppose firms' productivity bias can be expressed as

$$\iota_{it} = \Lambda\left(X_{it-1}; \lambda\right)$$

where X is a vector of observable firm attributes. Using productivity forecast errors, one can estimate the λ parameters of the productivity bias function and thereby recover consistent estimates of θ. Again, this leads to an additional estimation loop outside the main NPR estimation routine and the NPR algorithm becomes

1. Pick an initial vector of bias function parameters, $\hat{\lambda}_0$.
2. For iteration r, implement the NPR estimation algorithm using $\left(\mathbb{E}_{it}[y_{it+1}|I_{it}] - \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) - \Lambda(X_{it}; \hat{\lambda}_{r-1})\right)$ as the argument of the smooth function Ψ, to obtain coefficient estimates $\hat{\beta}_r$.
3. Use $\hat{\beta}_r$ to calculate $\mathrm{a}_{it} = \left(\mathbb{E}_{it-1}[y_{it}|I_{it-1}] - \int f(k_{it}, l_{it}; \hat{\beta}_r) dF_{it-1}(l_{it})\right)$ and $\mathrm{b}_{it} = y_{it} - f(k_{it}, l_{it}; \hat{\beta}_r)$ and their difference for all periods for which both expectations (needed to calculate a), and outcomes (needed to calculate b), are observed.
4. Obtain an updated estimate of the bias function parameters $\hat{\lambda}_r$ via estimation of

$$(\mathrm{a}_{it} - \mathrm{b}_{it}) = \Lambda\left(X_{it-1}; \lambda\right)$$

5. Compare $\hat{\lambda}_r$ to $\hat{\lambda}_{r-1}$. If the difference is sufficiently small the stop, if not then repeat from step 2.

In this case, the productivity forecast errors are given as

$$\begin{aligned}(\mathrm{a}_{it} - \mathrm{b}_{it}) &= (g(\omega_{it-1}) + \iota_{it}) - (g(\omega_{it-1}) + \xi_{it} + \epsilon_{it}) \\ &= \iota_{it} - \xi_{it} - \epsilon_{it} \\ &= \Lambda\left(X_{it-1}; \lambda\right) - \xi_{it} - \epsilon_{it}\end{aligned}$$

and since ξ and ϵ are each assumed i.i.d. and mean-zero, one can use the $(\mathrm{a}_{it} - \mathrm{b}_{it})$ terms to estimate the parameters of the productivity bias function Λ.

An alternative possibility is that firms' productivity bias is a function of their previous productivity. This is analogous to the 'excess sensitivity' or 'over-extrapolation' of output expectations discussed in literature on managerial and financial expectations (e.g.,

Barrero 2022). In our context, this type of bias means firms expect their productivity to increase/decrease between the current and next period if it has increased/decreased in recent periods. This type of bias is hard to accommodate if it specifically relates to firms' persistent productivity (i.e. ω_{it}). If, however, firms' biased beliefs over next-period productivity bias is determined by their current and recent *overall* productivity (i.e. $\omega_{it} + \epsilon_{it}$), one could modify the algorithm above accordingly.

1. Pick an initial vector of bias function parameters, $\hat{\lambda}_0$, and an initial vector of productivity terms (of length $N \times T$), $\hat{\varepsilon}_0$.

2. For iteration r, implement the NPR estimation algorithm using $\left(\mathbb{E}_{it}[y_{it+1}|I_{it}] - \int f(k_{it+1}, l_{it+1}; \beta) dF_{it}(l_{it+1}) - \Lambda(X_{it}; \hat{\lambda}_{r-1})\right)$ as the argument of the smooth function Ψ, to obtain coefficient estimates $\hat{\beta}_r$, where X_{it} includes $\hat{\varepsilon}_{it,r}$, $\hat{\varepsilon}_{it-1,r}$ and any other firm attributes that are believed to determine bias in firms' expectations of their next-period productivity.

3. Use $\hat{\beta}_r$ to calculate $\mathrm{a}_{it} = \left(\mathbb{E}_{it-1}[y_{it}|I_{it-1}] - \int f(k_{it}, l_{it}; \hat{\beta}_r) dF_{it-1}(l_{it})\right)$ and $\mathrm{b}_{it} = y_{it} - f(k_{it}, l_{it}; \hat{\beta}_r)$. Calculate b) for all periods that outcomes are observed and the difference (a-b) for all periods that both expectations (needed to calculate a), and outcomes (needed to calculate b), are observed.

4. Obtain an updated estimate of the bias function parameters $\hat{\lambda}_r$ via estimation of

$$(\mathrm{a}_{it} - \mathrm{b}_{it}) = \Lambda\left(X_{it-1}; \lambda\right)$$

   Obtain an updated estimate of the productivity terms as b)

$$\hat{\varepsilon}_r = y_{it} - f(k_{it}, l_{it}; \hat{\beta}_r)$$

5. Compare $\hat{\lambda}_r$ to $\hat{\lambda}_{r-1}$ and $\hat{\varepsilon}_r$ to $\hat{\varepsilon}_{r-1}$. If the difference across both vectors is sufficiently small the stop, if not then repeat from step 2.

Case 4: Imperfect Knowledge of the Production Technology

In subsection 2.5 we show that, if firms' have imperfect knowledge of the production function parameters then, in the case of Cobb-Douglas technology,

$$g(\omega_{it}) = \mathbb{E}_{it}[y_{it+1}|I_{it}] - (\beta_0 + \beta_k k_{it+1} + \beta_l \mathbb{E}_{it}[l_{it+1}|I_{it}]) - \Lambda\left(k_{it+1}, \mathbb{E}_{it}[l_{it+1}|I_{it}]; \Upsilon_i\right).$$

This is very similar to case 3 above, where bias in firms' expected productivity is a function of observables. As in this case, the baseline NPR algorithm can be embedded within an outer iteration loop to recover estimates of the Υ_i imperfect information terms. A nuance is that in this case, and assuming imperfect knowledge of the production function parameters varies across firms, the function $\Lambda(\cdot)$ should be estimated as a random coefficients model.

The exposition here assumes Cobb-Douglas technology but extends to any production function where imperfect knowledge of the functions' parameters can be rewritten as the function evaluated at the true parameter values and an additive term capturing firms' imperfect knowledge. However, we note that this is a particular case of imperfect information that confines firms' imperfect knowledge to the parameters of the production function rather than its functional form. Imperfect information of the latter type is generally insurmountable.

E Monte Carlo setup

Our Monte Carlo setup closely follows that of Ackerberg et al. (2015), the details of which are described in their Appendix. We repeat the key details here to ease readers' understanding of our results.

E.1 Production function and productivity process

We consider a gross output production function that is Leontief in the material input:

$$Y_{it} = \min\left\{\beta_0 K_{it}^{\beta_k} L_{it}^{\beta_l} e^{\omega_{it}}, \beta_m M_{it}\right\} e^{\epsilon_{it}},$$

where $\beta_0 = 1$, $\beta_k = 0.4$, $\beta_l = 0.6$ and , $\beta_m = 1$. ϵ_{it} is a mean-zero measurement error distributed i.i.d. over firms and time with standard deviation 0.1. The productivity shock ω_{it} follows and AR(1) process:

$$\omega_{it} = \rho\omega_{it-1} + \xi_{it},$$

with $\rho = 0.7$. The standard deviation of ω_{it} is constant over time and equal to 0.3, which is achieved via the parameterization of the normally distributed innovation ξ_{it} and the initial values ω_{i0}.

E.2 Firm choices

Firms' labor and material input choices are static in the sense that a firm's choice in period t has no implications for their choices in subsequent periods. This implies there are no adjustment costs in labor or materials which, combined with the other assumptions of our Monte Carlo setup, yield analytic solutions to the firms' decision problem.[46]

ACF consider several DGPs that differ in the timing of firms' labor choice. Specifically, ACF consider DGPs where firms' labor input in period t is chosen in an intermediate period $(t-1+b)$ with $0 < b < 1$, without full knowledge of the productivity shock ω_{it}. ACF include this dimension as their estimator is robust to such input timing, whereas the LP estimator is not. While our Monte Carlo algorithm can accommodate such timing, and we indeed replicate ACF's analysis below, we do not focus on this DGP in the results reported in the main text. Our baseline DGP therefore features firms that choose labor and materials concurrently in period t, with full knowledge of current-period productivity ω_{it}. Similarly, while ACF consider DGPs featuring firm-specific wages, we do not focus on them in our simulations with firms instead facing a common time-invariant wage.

Under these assumptions, the labor choice that maximizes firms' expected profits is

$$L_{it}^{*} = \left(\beta_0 \beta_l K_{it}^{\beta_k} e^{\omega_{it}}\right)^{1/(1-\beta_l)}, \tag{27}$$

and the optimal level of materials is

$$M_{it}^{*} = \beta_0 K_{it}^{\beta_k} L_{it}^{*\beta_l} e^{\omega_{it}}. \tag{28}$$

Capital is a dynamic input that evolves according to

$$K_{it} = (1-\delta)K_{it-1} + I_{it-1},$$

where $\delta = 0.2$ represents depreciation. Investment is subject to convex, firm-specific adjustment costs given by

$$c_i(I_{it}) = \frac{\phi_i}{2} I_{it}^2,$$

where $1/\phi_i$ is distributed log-normally across firms with a standard deviation (σ) of 0.6. While such firm-specific heterogeneity in capital adjustment costs renders the OP estimator inconsistent in all DGPs, it is included to generate variation in capital across firms comparable to that observed in both ACF's and our data.

46. Alternatively one could incorporate adjustment costs and solve the firms' problem numerically as in Bond and Söderbom (2005).

Under the assumptions described above, ACF extend the work of Syverson (2001) and Van Biesebroeck (2007) to obtain an analytical solution for optimal investment using Euler equation techniques:

$$\begin{aligned} I_{it}^{*} =& \frac{\beta}{\phi_i} \sum_{\tau=1}^{\infty} \left\{ (\beta(1-\delta))^{\tau} \left(\frac{\beta_k}{1-\beta_l} \right) \beta_0^{1/(1-\beta_l)} \right. \\ & \left. \times \exp \left[\left(\frac{1}{1-\beta_l} \right) \rho^{\tau+1} \omega_{it} + \frac{1}{2} \left(\frac{1}{1-\beta_l} \right)^2 \sum_{s=0}^{\tau} \rho^2 (\tau - s) \sigma_{\xi}^2 \right] \right\}, \end{aligned} \tag{29}$$

where $\beta = 0.95$ is the discount factor.[47]

As emphasized in the main text, the NPR estimator is robust to deviations from invertibility in either labor, materials or investment. ACF highlight the consistency of the LP estimator requires optimization error in labor but is undermined by optimization error in any other firm choices. The consistency of the ACF estimator under multiple types of optimization error is unclear. The DGPs examined in our Monte Carlo simulations therefore include optimization error in one or more inputs, which we model as

$$X_{it} = X_{it}^{*} e^{\xi_{it}^{x}},$$

for $X \in [L, I, M]$ where ξ_{it}^{x} is normally-distributed i.i.d. optimization error with mean zero and standard deviation $\sigma_{\xi^L}^2 = \sigma_{\xi^I}^2 = 0.37$ for labor and investment and $\sigma_{\xi^M}^2 = 0.185$ for intermediates (see footnote 22).

A detail to note is that optimal investment will have a different analytical solution from equation (29) when labor deviates from optimality. Specifically, when $\sigma_{\xi^L}^2 > 0$, ACF derive that optimal investment becomes

$$\begin{aligned} I_{it}^{*} =& \frac{\beta}{\phi_i} \sum_{\tau=1}^{\infty} \left\{ (\beta(1-\delta))^{\tau} \left(\frac{\beta_k}{1-\beta_l} \right) \beta_0^{1/(1-\beta_l)} \right. \\ & \left. \times \exp \left[\left(\frac{1}{1-\beta_l} \right) \rho^{\tau+1} \omega_{it} + \frac{1}{2} \left(\frac{1}{1-\beta_l} \right)^2 \sum_{s=0}^{\tau} \rho^2 (\tau - s) \sigma_{\xi}^2 \right] \right\} \\ & \times [\beta_l^{\beta_l/1-\beta_l} e^{(1/2)\beta_l^2 \sigma_{\xi^L}^2} - \beta_l^{1/1-\beta_l} e^{(1/2)\sigma_{\xi^L}^2}]. \end{aligned}$$

47. The assumption of constant returns to scale means that optimal investment does not depend on the current capital stock.

E.3 Expectations

The NPR estimator applied to a Cobb-Douglas production function requires data on one-period-ahead expectations of log labor and log output. To simulate these expectations, we rely on the expressions for firm choices described above.[48] Specifically, we have

$$\mathbb{E}_{it}[\omega_{it+1}|\Omega_{it}]] = \rho\omega_{it},$$

and hence

$$\begin{aligned}
\mathbb{E}_{it}\left[\ln\left(L_{it+1}\right)|\Omega_{it}\right] &= \mathbb{E}_{it}\left[\ln\left(L^*_{it+1}e^{\xi^L_{it+1}}\right)\right] \\
&= \mathbb{E}_{it}\left[\ln\left(L^*_{it+1}\right)\right] + \mathbb{E}_{it}\left[\xi^L_{it+1}\right] \\
&= \mathbb{E}_{it}\left[\ln\left(\left(\beta_0\beta_l K^{\beta_k}_{it+1}e^{\omega_{it+1}}\right)^{1/(1-\beta_l)}\right)\right] + 0 \\
&= \tfrac{1}{(1-\beta_l)}\left(\ln\left(\beta_0\beta_l\right) + \beta_k k_{it+1} + \mathbb{E}_{it}\left[\omega_{it+1}\right]\right) \\
&= \tfrac{1}{(1-\beta_l)}\left(\ln\left(\beta_0\beta_l\right) + \beta_k k_{it+1} + \rho\omega_{it}\right).
\end{aligned}$$

This in turn implies

$$\begin{aligned}
\mathbb{E}_{it}\left[\ln\left(Y_{it+1}\right)|\Omega_{it}\right] &= \mathbb{E}_{it}\left[\ln\left(\min\left\{\beta_0 K^{\beta_k}_{it+1}L^{\beta_l}_{it+1}e^{\omega_{it+1}}, \beta_m M_{it+1}\right\}e^{\epsilon_{it+1}}\right)|\Omega_{it}\right] \\
&= \mathbb{E}_{it}\left[\min\left\{\ln\left(\beta_0 K^{\beta_k}_{it+1}L^{\beta_l}_{it+1}e^{\omega_{it+1}}\right), \ln\left(\beta_m M_{it+1}\right)\right\}|\Omega_{it}\right] + \mathbb{E}_{it}\left[\epsilon_{it+1}|\Omega_{it}\right] \\
&= \mathbb{E}_{it}\left[\min\left\{\ln\left(\beta_0 K^{\beta_k}_{it+1}L^{\beta_l}_{it+1}e^{\omega_{it+1}}\right), \ln\left(\beta_m\beta_0 K^{\beta_k}_{it+1}L^{\beta_l}_{it+1}e^{\omega_{it+1}}e^{\xi^M_{it+1}}\right)\right\}|\Omega_{it}\right],
\end{aligned}$$

where the replacement of $\ln\left(\min\left\{a,b\right\}\right)$ with $\min\left\{\ln(a),\ln(b)\right\}$ that occurs in the transition from the first to the second equality implicitly assumes the arguments of the min function are always strictly positive. To evaluate the right hand side expectation we leverage results in Nadarajah and Kotz (2008), specifically their equation (11). This states that for $Y = \min(X_1, X_2)$, where (X_1, X_2) is a bivariate Gaussian random vector with means (μ_1, μ_2), variances (σ_1^2, σ_2^2) and correlation coefficient ρ, we have:

$$\mathbb{E}\left[Y\right] = \mu_1\Phi\left(\frac{\mu_2-\mu_1}{\theta}\right) + \mu_2\Phi\left(\frac{\mu_1-\mu_2}{\theta}\right) - \theta\phi\left(\frac{\mu_2-\mu_1}{\theta}\right), \tag{30}$$

48. The derivations here assume that firms take optimization error into account when forming expectations of next-period labor. However, since the optimization error becomes additive in the equation for log labor and is mean-zero, the expression is the same if one instead to suppose that firms are 'naive' in the sense that they continually expected optimal choices to be realised despite repeated experience to the contrary.

where $\theta = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$. Take then $X_1 \equiv \beta_0 + \beta_k k_{it+1} + \beta_l l_{it+1} + \omega_{it+1}$ and $X_2 \equiv \beta_m + m_{it+1}$. Conditional on Ω_{it}, (X_1, X_2) is bivariate normal with $\mu_1 = \mu_2 \equiv \mu = \mathbb{E}\left[\ln\left(\beta_0 K_{it+1}^{\beta_k} L_{it+1}^{\beta_l} e^{\omega_{it+1}}\right)|\Omega_{it}\right] = \mathbb{E}[\beta_m + m_{it+1}|\Omega_{it}]$ since optimality for the Leontief production function implies that $\ln\left(\beta_0 K_{it+1}^{\beta_k} L_{it+1}^{\beta_l} e^{\omega_{it+1}}\right) = \beta_m + m_{it+1}$ and the (log-) optimization error in intermediates has mean zero. Using equations (27) and (28) we can also obtain that $\sigma_1^2 = [\beta_L^2/(1-\beta_L)^2 + 1]\sigma_\xi^2$, $\sigma_2^2 = \sigma_1^2 + \sigma_{\xi^M}^2$ and $\rho\sigma_1\sigma_2 = \sigma_1^2$. This means that $\theta = \sqrt{\sigma_1^2 + \sigma_1^2 + \sigma_{\xi^M}^2 - 2\sigma_1^2} = \sigma_{\xi^M}$. Applying equation (30) to our context yields

$$\begin{aligned}
\mathbb{E}_{it}\left[\ln\left(Y_{it+1}\right)|\Omega_{it}\right] &= \mathbb{E}_{it}\left[\min\left(\beta_0 + \beta_k k_{it+1} + \beta_l l_{it+1} + \omega_{it+1}, \beta_m + m_{it+1}\right)|\Omega_{it}\right] \\
&= \mu\Phi\left(0\right) + \mu\Phi\left(0\right) - \sigma_{\xi^M}\phi\left(0\right) \\
&\approx 0.5\mu + 0.5\mu - 0.3989\sigma_{\xi^M} \\
&= \mu - 0.3989\sigma_{\xi^M}.
\end{aligned}$$

E.4 Simulation details

We consider a panel of 1000 firms observed over 10 time periods. For comparability with ACF's setup, our baseline specification follows their parameterization, which is intended to match key aspects of their data: 95% of the variation in capital being across-firm (versus within-firm), and the R^2 of a regression of capital on labor approximately equal 0.5. These moments are similar in our data at 93% and 0.4 respectively.

To avoid results depending on the arbitrary initial distribution of capital across firms, we initialize all firms with $k_{i0} = e^{-10} \approx 0$. We simulate firms for 100 periods and select the 10 periods in our estimation panel as the last of these, by which time the impact of the initial values appeared minimal.

E.5 Replication of ACF simulations

To confirm correct implementation of the MC routine, Table 11 repeats key DGPs of ACF's Table 1. The table shows moments of parameter estimates obtained by applying the OP, LP, ACF and NPR estimators to data simulated under the three distinct DGPs considered in Ackerberg et al. (2015). These are: DGP 1 which has serial correlation in wages and labor chosen at $t-b$, DGP 2 which has optimization in labor and DGP 3 which combines both DGP 1 and DGP 2. The coefficients obtained from an OLS levels regression of y_{it} on a constant, l_{it} and k_{it} are also provided for comparison. Each panel of the table contains results pertaining to the DGP indicated in the panel heading, while the results for the various estimators are given in rows. Within each DGP we show results

where materials has no measurement error (Panels A, C and E) and where materials do have measurement error (B, D and F). As acknowledged by ACF, the consistency of their estimator is compromised by the existence of multiple minima in their optimization routine, which causes it to sometimes converge to implausible parameter values (see footnote 16 in ACF). For this reason, we also provide results for the ACF estimator conditioning on simulation runs that obtained capital and labor parameter estimates within the zero to one range.

As expected, the OLS estimates and OP results are biased throughout, with bias in the latter due to firm-specific capital adjustment costs violating OP's monotonicity condition. As shown by ACF, the LP estimator is sensitive to assumptions regarding the timing of firms' labor decision and the presence of firm-specific wages: whereas LP outperforms ACF when labor is chosen at time t, wages are homogeneous across firms and there is optimization error in labor ('DGP 2' in the table), it fails to recover consistent estimates when labor is chosen at some intermediate '$t - b$' period and/or wages are heterogeneous across firms. The first sub-panel of within each DGP panel shows the ACF estimator is robust to DGPs 1 and 2 (if one discards 'implausible' results where either β_k or β_l lie outside the 0-1 range), but becomes biased when the assumptions of the first two DGPs occur in combination (DGP 3). The second sub-panels within each DGP show the performance of both the LP and ACF estimators deteriorate as measurement error is added to firms' material input. The NPR estimator, by contrast, achieves highly precise and consistent estimates across all DGPs and is unaffected by measurement error in materials.

Table 11: ACF Monte Carlo Results

	$\beta_l = 0.6$				$\beta_k = 0.4$				
	Mean	Median	S.D.	MSE	Mean	Median	S.D.	MSE	N runs
	DGP 1 - Serially Correlated Wages and Labor Set at Time $t-b$								
	A. No materials measurement error								
NPR	0.600	0.600	0.002	0.000	0.397	0.400	0.013	0.000	500
OLS	1.028	1.028	0.005	0.183	-0.010	-0.010	0.008	0.168	500
OP	0.865	0.865	0.006	0.070	0.177	0.177	0.011	0.050	500
LP	-0.000	-0.000	0.003	0.360	1.000	1.000	0.000	0.360	500
ACF	0.819	0.600	0.817	0.714	0.123	0.398	1.031	1.138	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.599	0.599	0.010	0.000	0.401	0.401	0.022	0.000	466
	B. Materials measurement error $\sim N(0, 0.5)$								
NPR	0.600	0.600	0.002	0.000	0.397	0.400	0.013	0.000	500
OLS	1.028	1.028	0.005	0.183	-0.010	-0.010	0.008	0.168	500
OP	0.865	0.865	0.006	0.070	0.177	0.177	0.011	0.050	500
LP	0.601	0.601	0.009	0.000	0.572	0.573	0.018	0.030	500
ACF	0.837	0.660	0.530	0.337	0.197	0.421	0.674	0.495	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.659	0.659	0.010	0.004	0.424	0.424	0.017	0.001	449
	DGP 2 - Optimization Error in Labor								
	C. No materials measurement error								
NPR	0.600	0.600	0.003	0.000	0.400	0.400	0.005	0.000	500
OLS	0.919	0.919	0.002	0.102	0.098	0.098	0.004	0.091	500
OP	0.840	0.840	0.004	0.057	0.162	0.161	0.008	0.057	500
LP	0.600	0.600	0.003	0.000	0.401	0.400	0.013	0.000	500
ACF	0.823	0.601	1.043	1.135	0.161	0.401	1.125	1.320	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.600	0.600	0.009	0.000	0.401	0.401	0.016	0.000	478
	D. Materials measurement error $\sim N(0, 0.5)$								
NPR	-2.4e+08	0.979	7.6e+08	6.4e+17	2.4e+08	0.027	7.6e+08	6.4e+17	500
OLS	1.000	1.000	0.001	0.160	0.000	0.000	0.002	0.160	500
OP	1.000	1.000	0.002	0.160	0.007	0.003	0.009	0.155	500
LP	1.000	1.000	0.004	0.160	0.001	0.001	0.013	0.160	500
ACF	-691.895	0.397	1151.381	1.8e+06	693.350	0.619	1149.967	1.8e+06	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.678	0.694	0.134	0.024	0.323	0.306	0.137	0.025	262
	DGP 3 - Optimization Error in Labor and Serially Correlated Wages and Labour Set at Time $t-b$ (DGP 1 plus DGP 2)								
	E. No materials measurement error								
NPR	0.600	0.600	0.002	0.000	0.399	0.400	0.009	0.000	500
OLS	0.948	0.948	0.005	0.121	0.100	0.101	0.009	0.090	500
OP	0.805	0.805	0.006	0.042	0.241	0.242	0.012	0.025	500
LP	0.372	0.372	0.004	0.052	0.813	0.814	0.020	0.171	500
ACF	0.632	0.608	0.239	0.058	0.343	0.374	0.306	0.097	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.608	0.608	0.007	0.000	0.372	0.374	0.022	0.001	491
	F. Materials measurement error $\sim N(0, 0.5)$								
NPR	0.600	0.600	0.002	0.000	0.397	0.400	0.013	0.000	500
OLS	1.028	1.028	0.005	0.183	-0.010	-0.010	0.008	0.168	500
OP	0.865	0.865	0.006	0.070	0.177	0.177	0.011	0.050	500
LP	0.601	0.601	0.009	0.000	0.572	0.573	0.018	0.030	500
ACF	0.811	0.660	0.491	0.285	0.229	0.422	0.627	0.422	500
ACF $\|(\beta_l, \beta_k \in (0,1))$	0.659	0.659	0.010	0.004	0.423	0.424	0.017	0.001	456

Note: number of replications given in the 'N runs' column pertains to the number of MC replications out of 500 that returned estimates with both β_l and β_k in the 0-1 range. The true values of β_l and β_k are 0.6 and 0.4 respectively. l is ln(labor), m is ln(materials), k is ln(capital) and i is ln(investment).

F Supplementary Data and Results

F.1 Expectations Survey Data

As noted in the main text the MES has been extensively used by researchers, mainly to examine the questions on management practices. The full surveys are available. The description and documentation of the full surveys are available from the ONS UK Data Service: https://datacatalogue.ukdataservice.ac.uk/studies/study/8557#documentation. For example, the specific 2017 survey is available for download at: https://doc.ukdataservice.ac.uk/doc/8557/mrdoc/pdf/8557_mes_2016_questionnaire_production.pdf.

We reproduce the key expectations questions for turnover in Figure 4 below. The format is identical for other inputs such as employment, investment and intermediate inputs. As noted in the main text, 2017 MES was conducted by paper, but 2020 was wholly online.

Figure 4: MES 2017 Turnover Questions

21. What are the dates of the 12 month period that you will be reporting for?

- Your reporting period should cover the calendar year 2016.
- If no figures are available for that period, your return should relate to a business year that ends between **6 April 2016 and 5 April 2017**.
- If you traded for only part of the year, please provide figures for the period in which you were trading.

From: D D M M Y Y Y Y 11 To: D D M M Y Y Y Y 12

22. For calendar years 2016 and 2017, what are the approximate pound sterling values of turnover, including exports and other receipts within this business? If applicable exclude freight charges, excise taxes and value added tax.

For 2016 calendar year..... £ 000

Forecast for 2017 calendar year..... £ 000

23. Looking ahead to the 2018 calendar year, what is the approximate pound sterling value of turnover you would anticipate for this business in the following scenarios, and what likelihood do you assign to each scenario?

Please refer to **Example B** if needed which explains how this question should be completed.

2018 scenarios, from lowest to highest	Approximate turnover in pounds (£) in 2018	Percentage likelihood (values in this column should sum to 100)
LOWEST	£ 000	%
LOW	£ 000	%
MEDIUM	£ 000	%
HIGH	£ 000	%
HIGHEST	£ 000	%
	Total	100 %

SPECIMEN

F.2 Translog production function

The parameters of a translog production function can be estimated via the basic NPR algorithm outlined in section 2 by setting

$$\begin{aligned} y_{it} &= f(k_{it}, l_{it}; \beta) + \omega_{it} + \epsilon_{it} \\ &= \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \beta_{k2} k_{it}^2 + \beta_{l2} l_{it}^2 + \beta_{kl} k_{it} l_{it} + \omega_{it} + \epsilon_{it}. \end{aligned}$$

In this case, construction of the argument of the non-linear function NPR uses to control for ω requires data on $\mathbb{E}[l_{it+1}^2|\Omega_{it}]$. We calculate this as

$$\begin{aligned} \mathbb{E}[l_{it+1}^2|\Omega_{it}] &= \text{Var}[l_{it+1}|\Omega_{it}] + \mathbb{E}[l_{it+1}|\Omega_{it}]^2 \\ &= \sigma_l + \mu_l^2, \end{aligned}$$

where μ_l and σ_l are the mean and variance of the lognormal distributions we fit to firms' subjective expectations of year-ahead labor. Our assumptions on capital mean $k_{it+1} \in \Omega_{it}$, and hence the other additional terms in the control function are straightforward to obtain as k_{it+1}^2 and $k_{it+1}\mathbb{E}[l_{it+1}|\Omega_{it}]$.

Tables 12-14 contain translog parameter estimates obtained via NPR, ACF and various OLS estimators. The $\bar{l}$ and $\bar{k}$ rows contain the estimation sample means of l and k respectively, which are used to calculate the mean partial derivatives given in the $\frac{\partial y}{\partial l}$ and $\frac{\partial y}{\partial k}$ rows. The 'Cobb-Douglas Wald' row gives the Wald statistic of a test that the non-Cobb-Douglas parameters $(\beta_{l2}, \beta_{k2}, \beta_{lk})$ are all equal to zero and Row 'Cobb-Douglas P-Value' gives the corresponding p-value. These rows show the NPR estimates fail to reject Cobb-Douglas technology in the electronics industries. Although Cobb-Douglas technology is rejected by NPR in the retail industry, the mean partial derivatives of 0.91 with respect to labor is within the 95% confidence interval of the NPR Cobb-Douglas estimate of β_l (0.80 with standard error 0.11), while the mean partial derivative with respect to capital is identical to the NPR Cobb-Douglas estimate of β_k (see Table 8). We therefore focus on the Cobb-Douglas results in our main analysis.

Table 12: Electronics: Translog Production Function Coefficient Estimates

	NPR	OLS	OLS FD	OLS FE	ACF
β_{l1}	0.95*	0.97***	0.63**	0.83***	0.91***
	(0.55)	(0.17)	(0.28)	(0.27)	(0.24)
β_{k1}	-0.04	0.04	0.12	0.24*	-0.21
	(0.39)	(0.09)	(0.09)	(0.13)	(0.24)
β_{l2}	0.04	0.07**	0.00	-0.00	0.06
	(0.09)	(0.03)	(0.04)	(0.04)	(0.11)
β_{k2}	0.05**	0.06***	0.01	0.00	0.05
	(0.02)	(0.01)	(0.01)	(0.01)	(0.04)
β_{lk}	-0.08	-0.12***	-0.03	-0.05*	-0.07
	(0.07)	(0.03)	(0.02)	(0.03)	(0.09)
$\bar{l}$	3.95	3.97	3.98	3.98	3.97
$\bar{k}$	7.60	7.61	7.63	7.64	7.61
$\frac{\partial y}{\partial l}$	0.70***	0.60***	0.43***	0.46***	0.81
	(0.11)	(0.04)	(0.13)	(0.09)	(0.81)
$\frac{\partial y}{\partial k}$	0.39***	0.44***	0.09	0.07	0.25
	(0.09)	(0.03)	(0.06)	(0.06)	(0.19)
Cobb-Douglas Wald	4.92	52.42	2.18	5.25	2.33
Cobb-Douglas P-Value	0.18	0.00	0.54	0.15	0.51
N obs.	472	917	458	895	917
N firms	422	422	411	400	422

Note: dependent variable is log turnover. Parentheses contain standard errors. NPR standard errors calculated from 100 bootstrap replications. Column titles indicate estimation methods. */**/*** denote significance at the 10/5/1 percent level respectively.

Table 13: Retail: Translog Production Function Coefficient Estimates

	NPR	OLS	OLS FD	OLS FE	ACF
β_{l1}	1.77***	1.15***	0.73**	0.60**	1.10
	(0.35)	(0.09)	(0.35)	(0.26)	(1.90)
β_{k1}	0.09	0.22***	0.02	0.02	-0.02
	(0.18)	(0.07)	(0.04)	(0.04)	(0.32)
β_{l2}	-0.08**	-0.02	-0.03	0.00	-0.35
	(0.04)	(0.02)	(0.05)	(0.04)	(0.41)
β_{k2}	0.01	0.01**	0.00	0.00	0.02
	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)
β_{lk}	-0.03	-0.03*	-0.02	-0.01	-0.08
	(0.04)	(0.02)	(0.02)	(0.02)	(0.07)
$\bar{l}$	3.91	3.92	3.93	3.93	3.92
$\bar{k}$	6.80	6.69	6.83	6.71	6.69
$\frac{\partial y}{\partial l}$	0.91***	0.79***	0.43***	0.52***	-2.11
	(0.11)	(0.04)	(0.09)	(0.08)	(2.05)
$\frac{\partial y}{\partial k}$	0.16***	0.26***	-0.01	-0.01	-0.01
	(0.05)	(0.03)	(0.02)	(0.02)	(0.04)
Cobb-Douglas Wald	9.44	24.95	2.07	1.10	2.50
Cobb-Douglas P-Value	0.02	0.00	0.56	0.78	0.47
N obs.	2084	4057	2023	3967	4057
N firms	1853	1853	1807	1763	1853

Note: dependent variable is log turnover. Parentheses contain standard errors. NPR standard errors calculated from 100 bootstrap replications. Column titles indicate estimation methods. */**/*** denote significance at the 10/5/1 percent level respectively.

Table 14: Restaurants: Translog Production Function Coefficient Estimates

	NPR	OLS	OLS FD	OLS FE	ACF
β_{l1}	0.66*	0.79***	0.33	0.20	1.07***
	(0.35)	(0.09)	(0.34)	(0.31)	(0.40)
β_{k1}	0.50*	0.06	-0.10	-0.16	0.17
	(0.28)	(0.09)	(0.27)	(0.22)	(0.45)
β_{l2}	0.01	0.00	-0.00	0.01	-0.12
	(0.05)	(0.03)	(0.04)	(0.04)	(0.11)
β_{k2}	-0.02	0.01*	-0.00	-0.00	-0.03
	(0.04)	(0.01)	(0.01)	(0.01)	(0.04)
β_{lk}	0.01	-0.00	0.04	0.05	0.11
	(0.07)	(0.04)	(0.04)	(0.04)	(0.12)
$\bar{l}$	4.48	4.51	4.53	4.54	4.51
$\bar{k}$	7.67	7.69	7.74	7.73	7.69
$\frac{\partial y}{\partial l}$	0.80***	0.76***	0.63***	0.68***	0.90**
	(0.08)	(0.05)	(0.09)	(0.09)	(0.40)
$\frac{\partial y}{\partial k}$	0.28***	0.26***	-0.00	0.02	0.20
	(0.07)	(0.05)	(0.10)	(0.07)	(0.32)
Cobb-Douglas Wald	0.74	13.07	3.83	5.58	1.46
Cobb-Douglas P-Value	0.86	0.00	0.28	0.13	0.69
N obs.	462	879	421	838	879
N firms	430	430	392	389	430

Note: dependent variable is log turnover. Parentheses contain standard errors. NPR standard errors calculated from 100 bootstrap replications. Column titles indicate estimation methods. */**/*** denote significance at the 10/5/1 percent level respectively.