Estimating outcomes in the presence of endogeneity and measurement error with an application to R&D♣

Dakshina G. De Silva[†] Timothy P. Hubbard[‡] Anita R. Schiller[§] Mike G. Tsionas[**]

January 29, 2023

Abstract

We adopt a Bayesian econometric technique to address issues of endogeneity and measurement error when estimating outcomes while also tackling censoring. We motivate our study based on the theoretical framework laid out by Dasgupta and Stiglitz [1980] to highlight the endogeneity issue by investigating the relationship between market structure and innovation. We apply our method to estimate the R&D expenditures for Chinese manufacturing firms to highlight the importance of the econometric issues. Reduced-form results suggest a nonlinear relationship between market concentration and R&D expenditures, while our approach suggests a strictly positive relationship consistent with canonical theoretical models built on oligopolistic competition.

JEL Classification: C11, C13, O30.
Keywords: measurement error; endogeneity; empirical likelihood; Bayesian methods; Markov chain Monte Carlo; research and development.

♣ We greatly appreciate suggestions from the editor, Dan Bernhardt, two anonymous referees, Mons Chan, George Deltas, as well as participants at the 20th Annual International Industrial Organization Conference, the 97th Annual Western Economic Association International Conference, and the Korea Institute for International Economic Policy.
[†] Department of Economics, Lancaster University Management School, Lancaster LA1 4YX, UK and Faculty of Management, Social Sciences & Humanities, General Sir John Kotelawala Defence University, Kandawala Rd, Ratmalana, 10390, Sri Lanka; email: d.desilva@lancaster.ac.uk.
[‡]Department of Economics, Colby College, 5242 Mayflower Hill Drive, Waterville, ME 04901, USA; email: timothy.hubbard@colby.edu
[§] Department of Economics, Lancaster University Management School, Lancaster LA1 4YX, UK; email: anita.schiller@lancaster.ac.uk
[**] Corresponding author: Department of Economics, Lancaster University Management School, Lancaster LA1 4YX, UK; email: m.tsionas@lancaster.ac.uk

1. Introduction

Estimating outcomes using ordinary least squares (OLS) in the presence of endogeneity and/or measurement error can result in biased and inaccurate findings. For example, industrial economists have long considered the relationship between incentives to innovate and market structure in trying to determine market conditions that encourage research and development (R&D) activity. Understanding this relationship is important as R&D by firms leads to innovation, from which companies can obtain economic rents.[1] Schumpeter [1942] hypothesized that large firms in concentrated markets are likely to innovate. Many researchers have since studied this relationship—now known as the *Schumpeterian hypothesis*.[2]

One takeaway from this research is that there is still a need to account for issues of simultaneity between market structure and R&D. For example, Cohen [2010] noted "an area where this literature on the tie between R&D and firm size is relatively mute is the endogeneity of firm size with respect to R&D and innovation." Further, as mentioned by Aghion et al. [2005], accurately accounting for the number of competitors is also challenging and can lead to measurement error. Empiricists have documented (using Wu-Hausman tests) that, in such instances, orthogonality conditions required for OLS estimation do not hold.[3] In a recent paper, Li et al. [2021] also mention the bias in OLS estimators. This could be exacerbated when using a sample of survey data to construct variables to control for agglomeration or market concentration ratios where least squares estimation may be inappropriate and richer econometric strategies are needed.

Hence, in this paper, we adopt an alternative approach based on econometric techniques developed by Schennach [2005, 2014] that addresses endogeneity and measurement error while paying attention to censoring issues (for example, when firms do not invest in any R&D efforts). Thus, our econometric contribution is twofold in the sense that we (i) deal with measurement error in critical variables of the model, and (ii) deal with endogeneity—in this case, of rivals' R&D, market concentration, and the

[1] R&D efforts can focus on either reducing the cost of producing a product (R&D related to process innovations) or on improving the end-product, as well as introducing new products (R&D related to product innovations) by building directly upon existing products or by introducing new varieties.

[2] Cohen [2010] recapped empirical research that has considered hypotheses in the Schumpeterian tradition, updating his previous surveys (Cohen and Levin [1989] as well as Cohen [1995]). In addition, Gilbert [2006] surveyed literature specifically related to the Schumpeterian hypotheses around market structure and firm size while Ahuja et al. [2008] surveyed related research in the management literature.

[3] See, for example, Levin and Reiss [1984] as well as Levin et al. [1985].

number of rival firms. We use a coherent and unified framework by combining Entropic Latent Variable Integration via Simulation (ELVIS) (which deals with measurement error in nonlinear regressors) with Bayesian Exponentially Tilted Empirical Likelihood (BETEL) (which deals generically with moment conditions using instruments) which we organize in a general formulation that can, in turn, be estimated using fast Markov Chain Monte Carlo-based (MCMC) methods.

Considering the literature on R&D, the common approaches researchers have adopted—recognizing the potential simultaneity between innovative efforts and market concentration—include instrumenting for concentration or estimating multi-equation models that treat concentration and R&D as endogenous.[4] Blundell et al. [1999] proposed an approach for addressing the endogeneity issue concerning market structure and R&D by using lagged variables in panel data. They found that concentrated industries produce fewer innovations though, within an industry, larger firms generate more innovations. One challenge with this approach is that researchers may have access to limited data (for example, a panel of data might span only a couple of years) which might preclude the ability to use lagged variables.

We apply our estimator to Chinese firm-level data for which a short panel exists so constructing lagged variables removes important observations. The focus of our research is on inputs or efforts—R&D expenditures—and not on output or productivity of R&D (for example, patents per dollar of R&D spending). Our data is firm-level and firms are assigned to industries at the four-digit level under the Standard Industrial Classification (SIC) codes.[5] Reduced-form (OLS) results suggest a nonlinear relationship between measures of market concentration and R&D expenditures. This finding is consistent with those of the literature, though we apply this approach to a new data set.

Estimating analogous models using our hybrid results leads to similar estimates for a number of covariates that are plausibly exogenous. However, our hybrid Bayesian approach leads to very different results for the factors that we argue are endogenous or plagued by measurement error like measures of market concentration.

[4]As an example of the former, Levin et al. [1985] instrumented for market concentration (four-firm concentration ratio) in estimating the effect of market structure on R&D intensity. For the latter, Levin and Reiss [1984, 1988] as well as Connolly and Hirschey [1984] used a four-equation system specifying relationships between profits, R&D, advertising, and concentration, allowing for nonlinearities.

[5]As detailed in the survey by Cohen [2010], data collection on R&D efforts and output measures in the United States is often not disaggregated enough to consider. Instead, research has focused on data from Canada and Europe (for example, Community Innovation Survey data) primarily.

Specifically, our approach suggests a strictly positive relationship exists in our data—the more concentrated the industry, the more expenditures on R&D. A positive relationship is consistent with canonical theoretical models built on oligopolistic competition such as that of Dasgupta and Stiglitz [1980] and offers support of Schumpeter's hypothesis. Further, we follow our estimation results by shutting down elements of our approach and by performing MCMC exercises which encourage confidence in our approach when issues of endogeneity and measurement error are present in the data.

Our hybrid approach addresses a number of issues: censoring, measurement error in explanatory variables, and endogeneity. The latter two challenges are particularly important as measurement errors can result not only from poor data but when variables are constructed using a subset of all firms while endogeneity is prevalent in similar applications. For example, one firm's R&D expenditures might depend on the R&D expenditures of rival firms. However, that aggregate value can also be plagued by measurement error if all firms are not observed or if firms are not correctly classified into industries. This issue could be exacerbated, especially if they participate in multiple industries but variable construction is built around only their primary industry.

We acknowledge that endogeneity can be dealt with in a generalized method of moments (GMM) framework using moment conditions. However, there is increasing evidence that GMM can behave erratically in finite samples (see the initial work of Hansen et al. [1996] for example) and its behavior is not ideal when instruments are weak or invalid. Hence, our hybrid approach that is based on the BETEL of Schennach [2005] can be thought of as a Bayesian version of GMM which addresses the shortcomings of a GMM strategy. In this framework, although we cannot entirely solve the general problem of weak or invalid instruments, we have, at least, a principled way to test these assumptions. In our application, we use MCMC methods to provide access to the posterior implied by moment conditions, as suitable instruments are available.

Measurement errors complicate the analysis considerably. There is no standard approach to the problem, though Lewbel [1997] proposed one solution in which functions of the data can be used as instruments in multiple-stage least-squares regression using higher moments of the data. We use the ELVIS method of Schennach [2014] and specialized MCMC algorithms based on Girolami and Calderhead [2011] to access the posterior of the model. ELVIS deals explicitly with measurement error

problems and can be embedded into BETEL, allowing us to address these issues by employing an estimator (from the classical perspective) that has both good asymptotic and finite-sample properties.

Our paper is structured as follows: in Section 2, we appeal to a simple theoretical model which highlights the simultaneity between market structure and R&D investment choices. That model motivates our econometric concerns and allows us to demonstrate the inadequacy of OLS. In Section 3, we discuss firm-level panel data which we summarize, noting interesting R&D investment patterns and relationships. We initially ignore our econometric concerns by formally estimating empirical models in Section 4 using standard techniques including instrumental variables (IV) method to establish baseline results. Our hybrid Bayesian estimation strategy, meant to address the concerns around censoring, measurement error, and simultaneity, is articulated and estimated in Section 5 where we find important differences in the main takeaways. In Section 6, we present simulation evidence that allows us to compare our approach with nested models which include OLS as well as specifications which account for just measurement error or only endogeneity. Lastly, In Section 7, we summarize and conclude our research.

2. Conceptual framework

In this section, we simplify the classic Dasgupta and Stiglitz [1980] model based on Cournot competition, in which R&D expenditures are explicitly a part of firms' strategies. Consider an industry with n identical firms. Firm i chooses its output level $q_i \geq 0$ and an amount to spend on R&D, $x_i \geq 0$. R&D investments are costly but reduce the firm's unit cost of producing output. That is, firm i's marginal cost $c_i = c(x_i)$ where $c'(x_i) < 0$. Profit for firm i can be expressed as

$$\pi_i(q_i, x_i; Q_{-i}) = p(Q)q_i - c(x_i)q_i - x_i$$

where $Q = \sum_{i=1}^{n} q_i$ is aggregate output in the industry, $Q_{-i} = Q - q_i$ corresponds to the output of firm i's rivals which i takes as given, and $p(Q) = a - bQ$ is the industry demand given aggregate output.

A Nash equilibrium with free entry and exit $[n^*, (q_1, x_1), (q_2, x_2), \ldots, (q_{n^*}, x_{n^*})]$ ensures that no active firm has incentive to change its output or R&D investment decision and that no potential entrants (incumbents) have incentive to enter (exit) the industry given the decisions of other firms. Assuming

symmetry, the first-order conditions for profit maximization equate marginal cost of R&D expenditures with the maraginal benefit in the form of a reduced marginal cost of production

$$1 = -q^* c'(x^*) \tag{1}$$

and

$$\frac{p(Q^*) - c(x^*)}{p(Q^*)} = \frac{\partial p(Q^*)}{\partial q^*} \frac{q^*}{p(Q^*)} \tag{2}$$

which, letting $s \equiv q/Q$, can be transformed to yield the Lerner Index (LI)

$$\frac{p(Q^*) - c(x^*)}{p(Q^*)} = \frac{s^*}{\epsilon(Q^*)} \tag{3}$$

where $\epsilon(\cdot)$ is the price elasticity of demand.[6]

In a model with identical firms, $s^* = 1/n^*$, so equation (3) can be rewritten as

$$p(Q^*)\left(1 - \frac{1}{\epsilon(Q^*)n^*}\right) = c(x^*) \tag{4}$$

which makes clear that an increase in the number of firms reduces the output of each individual firm. However, the marginal benefit of R&D expenditures is proportional to the output level of an individual firm as reflected in condition (1). Thus, as the number of firms increases, each individual firm reduces output, which reduces the marginal benefit from R&D expenditures, meaning R&D spending falls for firms active in the industry. The model then suggests that individual firms in industries with more rivals will spend less on R&D; the lower the number of competitors (the more concentrated the market in our symmetric model), the more firms will individually spend on R&D. Moreover, aggregating across firms and using zero profit conditions means

$$\frac{n^* x^*}{p(Q^*)Q^*} = \frac{1}{n^* \epsilon(Q^*)} \tag{5}$$

which relates the amount of R&D spending in an industry to the share of industry sales.

An important insight of this model is that industrial concentration and research intensity are simultaneously determined—measures of market concentration and R&D efforts are endogenous. Firm investment decisions are simultaneous in this model, which means including variables capturing rivals' R&D efforts in empirical work will imply the variables are correlated with error terms.

[6]This model can be extended to allow for asymmetric firms which differ in their R&D investments and hence marginal costs, but the LI can still be derived from the structure of the model.

3. Data

In our application, we use yearly data from The Annual Survey of Industrial Firms (ASIF) conducted and maintained by China's National Bureau of Statistics (NBS) for the period of 2005 to 2007.[7] The data include firms with annual sales of at least five million renminbi (RMB) or approximately $735,000—this accounts for over 85 percent of Chinese industrial output. Firms must report their (unique) legal identification number as well as the name of the firm, allowing us to track firms that are observed in multiple years and exploit the panel structure of our data.

According to the classifications of the NBS, the data comprise three types of firms: (1) privately owned enterprises; (2) foreign multinationals operating in China; (3) state-owned enterprises. Further, firms are divided into three main categories of industries: mining, manufacturing, and production and distribution of electricity, gas and water. In our study, we concentrate only on the subsample of manufacturing firms during the years 2005–2007, which is when the ASIF included firm-specific R&D investments. Focusing our sample on this data yields 739,212 observations—specifically, we observe 217,653, 245,094, and 276,465 manufacturing firms during years 2005, 2006, and 2007 respectively. Lu and Tao [2009] noted that this upward trend in manufacturing firms is due to the rapid growth in manufacturing sectors during the sample period, increasing the number of firms with annual sales which exceed the five million RMB threshold for inclusion in the ASIF dataset.

In addition to information on the total expenditures on R&D, the data include information on production activities (employment, capital, intermediate inputs, sales), balance sheet statements (current and total assets, liabilities, inventories, financing costs, taxes paid, operating costs, profits), and firm characteristics (industry classification for primary and secondary products, location, ownership type). Moreover, the ASIF data contain firm-level trade data which allow us to distinguish exporters from non-exporters. We define the variables which we work with in Table A.1 and present pairwise correlations in Table A.2, in the Appendix A.

For each firm in the ASIF data set, the location information details its address as well as the name of city, district, and province where it is located. Naturally, the more precise information on a firm's location allows for more accurate construction of agglomeration and other geographic-based variables; see, for example, Rosenthal and

[7]These data were also used by Bai et al. [2006], Cai and Liu [2009], as well as Lu and Tao [2009].

Strange [2003]. We define geographic units to be at the 2010 zip-code level which is matched with each firm address in the data and identifies 31,046 areas where manufacturing firms are located.[8] We also have information on each firm's primary industry code at the four-digit SIC level, for which we see 525 different industries represented in our data. This is important as empirical researchers in this literature have found that industry fixed effects explain a large share of variance in R&D-related dependent variables; for example, see Scherer [1967] and Wilson [1977]. Geroski [1990] even found that his fundamental result of a positive relationship between competition and innovation was reversed if industry effects were not included.

In our approach, we go beyond this by controlling not just for industry effects but, because we have a panel of data, for firm-level fixed effects. Moreover, for a given firm, we construct industrial agglomeration measures not only at the zip code level, but within a given SIC code for each zip code for a given year. Note however, that while this is an important link between theoretical and empirical work, the potential for measurement error is now much greater for two reasons: (i) variable construction is at the industry level; (ii) only a subset of potential rivals are observed. The former stems from firms being classified in the data as participating in a primary industry, which is certainly convenient for empirical work, but also difficult to think about for multi-product firms who are then omitted from other industries in which they may be active.[9] The latter stems from smaller firms not being included in the data, and hence never included in variables concerning rivals' behavior that are computed from the raw data. We view both of these issues as important sources of measurement error.

In Figure 1 we map total R&D expenditures within each district in China in 2005 (Panel A) and 2007 (Panel B), respectively. Comparing the two maps demonstrates substantial growth in innovative efforts across these two years in our data. Both figures make clear that the bulk of R&D investments occur along the coast and eastern parts of China, consistent with where most of the economic activity occurs. The maps demonstrate both increased efforts within a number of districts that were already

[8]Due to growth in China, its administrative boundaries of cities, zip codes, counties, or even provinces have experienced changes in the last thirty years. We geo-code each address at the 2010 zip code level which is when boundaries are most disaggregated. Fixing these geographic units maintains the same area definitions during our analysis.

[9]One element of a dataset that could help address this concern from the raw data is if R&D expenditures were disaggregated and somehow associated with the various products a firm produces so that R&D investments could be "allocated" towards the different industries. This is not the case in our data nor most R&D-based datasets with which we are familiar.

active in R&D spending, as well as the proliferation of these efforts to new districts which see R&D investments from its constituent firms from 2005–2007. Still, a number of districts, particularly in central and western China see little R&D activity; these districts are often geographically much larger, but house few manufacturing firms. In Figure 2, we plot the locations of all firms in our data which meet the requirements discussed previously and are used in our forthcoming analysis.

In Table 1, we report summary statistics for important variables in our data. Specifically, we report the average and standard deviation (below and in parentheses) for select variables. There are 525 four-digit SIC industries within the manufacturing sector. Comprising these industries are 324,463 unique firms appearing at least once in our data. On average, these firms spend about 269.6 million RMB on R&D. These firms on average enjoy profits of 14,348 million RMB deriving from average sales of 82,856 million RMB. As in Aghion et al. [2005], we construct a LI to represent product market competition as follows:

$$\mathrm{LI}_{it} = \frac{operating\ profit_{it} - financial\ cost_{it}}{current\ sales_{it}}. \tag{6}$$

The LI is firm- and time-specific and has several advantages over indicators such as market share, firm concentration ratios, or the Herfindahl-Hirschman Index (HHI). These other measures rely more directly on precise definitions of geographic and product markets (Aghion et al. [2005]). While we know the SIC code that a firm operates in (allowing us to understand their output market), it can be difficult to define a specific geographic area relevant for their competition as more than 23% of firms operate in international markets (the dummy variable Exporter takes a value of one if the firm exports product outside of China, and zero otherwise). One challenge with constructing the LI via (9) is that some firms report negative operating profits in a given year. In Table 1, we provide summary statistics for both the relevant sample given considerations previously noted, as well as for the restricted sample in which we only consider observations with nonnegative operating profits so that the LI measure is properly characterized.

In Figure 3, we plot the log of R&D expenditures (using the right-hand axis) against the LI measure. The LI, which is bound between zero and one, takes on an average value of 0.202 and the median value is 0.046 conveying that many markets are

quite competitive, though the average is being brought up by a few very concentrated industries. This observation is made clear by a second depiction within this figure which plots a histogram over the LI measure to give a better sense of this distribution (using the left-hand axis). Given there is substantial mass near the boundaries, the confidence interval on the log of R&D expenditures is widest in the middle. The inverted-U shape depicted in Figure 3 is consistent with what others in the literature have found (see, for example, Levin et al. [1985]) and matches the trend that Aghion et al. [2005] observed at the industry level. The inverted-U relationship suggests that R&D expenditures are highest in modestly concentrated industries—on average, there are 2–3 firms in industries where the inverted-U obtains its maximum suggesting these industries can best be characterized by oligopoly.

Given the prevalence of firms with much lower LI, perhaps unsurprisingly, firms in our data compete in industries with higher competition—a firm faces on average about 5.8 rivals in our data. These firms have about 200 employees and have been in business for over eight years on average. About a quarter of all firms produce two outputs, but less than 10% produce three or more different products. In our analysis, we consider the firm to compete in the industry that corresponds with the four-digit SIC code of its primary output.[10]

With these data in mind, we apply empirical models that are consistent in spirit with past work in this literature and discuss estimation results in the next section. This allows us to establish some benchmark findings which can be contrasted with results from our hybrid Bayesian estimation strategy which we offer in Section 5.

4. Descriptive Regression Results

Given the trends presented in the previous section, we seek to evaluate the relationship between market power, competition, and R&D efforts as measured by expenditures. Specifically, we first model these relationships as

$$\begin{aligned} \log\left(\mathrm{R\&D}_{ijlt}\right) = {} & f\left(\mathrm{LI}_{ijt}\right) + g\left(\mathrm{competition}_{ijlt}\right) \\ & + \alpha \log\left(\mathrm{rivals'\ R\&D}_{ijlt}\right) + \beta X_{it} + \gamma W_{lt} \\ & + \delta_i + \tau_t + \varepsilon_{ijlt}. \end{aligned} \tag{7}$$

[10] In considering Chinese firms, one may wonder about the ownership of these enterprises. About 94% of firms in our data have no governmental ownership; the Chinese government has a minority stake in 1.2% of firms, and a majority stake in 4.8% of firms. We account for this structure in our empirical work going forward.

In these specifications, we recognize that the R&D investments of a given firm may depend on factors that are firm-, industry-, location-, and time-specific. We derived the LI from our theoretical model in equation (3) and presented its empirical counterpart to this measure in (6). We proxy for competition by counting the number of firms operating in the same industry, in the same geographic location (zip code), during the same year. We model $f(\cdot)$ and $g(\cdot)$ as cubic polynomials of their respective arguments to allow for richer relationships and given the inverted-U pattern observed in Figure 3.[11] As suggested by the theoretical model presented earlier, we also consider that firms may be investing in R&D strategically in equilibrium and so these decisions may depend on the R&D investments of rivals—firms operating in the same industry j as firm i within a given location l at time t. Industry j is specified at the four-digit SIC code, location l corresponds with a zip code, and time t to a year in the data.

The theory presented highlights that there are endogeneity concerns with respect to the regressors related to the Lerner Index (LI), number of rivals, and rivals' average R&D. These variables are simultaneously determined with a firm's own R&D expenditures in the motivating theoretical structure.

Beyond this, we include covariates that are either specific to a firm and observed to vary over time (for example, the number of employees at a firm, age of the firm, exporter status, type of ownership, whether the firm is a multi-product firm; these are contained in X_{it}) or to a given location at a point in time (for example, the number of national and provincial universities in a region; these are contained in W_{lt}).[12] These variables are exogenous regressors potentially important in explaining variance in firms' own R&D choices. In our models we also include firm (δ_i) and time (τ_t) fixed effects. The inclusion of firm fixed effects implies that identification of the model is driven by changes within a firm across years in the data. Since each firm is assigned to only one industry (that of their primary output) and industry does not change over our sample, these firm effects capture industry effects which are critical to account for. The term ε_{ijlt} represents an independently distributed error term that is firm-, rivals-,

[11] A cubic specification nests a quadratic one, thereby flexibly allowing for such a relationship to arise without imposing it. This practice is common when modeling potential inverse-U relationships; see, for example, Grossman and Krueger [1995], De Silva et al. [2016], and De Silva et al. [2021] who all adopt this approach in modeling the environmental Kuznets curve.

[12] Belderbos et al. [2021] also showed that proximity of central R&D units are important orchestrators of research collaboration with universities.

location-, and time-specific reflecting influences on (log) $R\&D_{ijlt}$ from factors not included in our model.[13] All of these variables are included in all models we estimate.

We present baseline estimates in Table 2, which provides OLS coefficient estimates from six different specifications nested by the empirical model presented in equation (7). In the first column, we focus primarily on the relationship between LI and R&D expenditures, controlling for a firm's rivals' spending as well as the firm's size, age, and exporter status. Given we represent the LI by a polynomial, interpreting a given coefficient directly is somewhat difficult. As such, we depict the estimates from column (6) in Figure 4, Panel A which confirms a nonlinear relationship with an interior peak. Given our log-log specification concerning rivals' average R&D expenditures, that coefficient can be interpreted as an elasticity which suggests that if the average of a firm's rivals' R&D spending increases by 1%, a firm's own R&D investments would increase 6.6% in column (1). The larger a firm, as measured by the number of employees, the more a firm invests in R&D—a 1% increase in employment corresponds with an 11.1% increase in R&D expenditures. Age is not significant but being an exporter means R&D spending is nearly 10% higher than non-exporters.

In the column (2), we present estimates from a model in which we omit our LI measure, but include a cubic relationship concerning the logarithm of the number of rivals in the market as a measure of competition. Plotting the estimates shows the relationship between the number of rivals and the logarithm of a firm's R&D spending is nearly linear—as the number of rivals increases by one firm, log(R&D) spending decreases by 0.0012. The effects of other covariates are nearly identical to those detailed above. In fact, the estimates in column (3) suggest inclusion of both our LI measure and the competition measure with the same set of covariates leaves our estimates, and their statistical significance, essentially unchanged. Pushing harder on this, if we include controls for multi-product firms and the number of higher education schools in a zip code (as a proxy for how many researchers may live in the area), all results remain, as reflected by the estimates in column (4). Given we consider Chinese data in which the government occasionally holds an ownership stake in some firms, we wanted to make sure that our results are robust to this structure. In column (5), we

[13] This is an important assumption needed for this approach. Of course, our focus of issues on endogeneity (via simultaneity) means there is correlation between ε_{ijlt} and those variables in our model which is likely aggravated by the measurement error concerns we have raised. Remember that we are estimating this model primarily to establish baseline results for comparison.

consider a model in which we include directly the share of government ownership in the firm. That covariate is not significant and our earlier discussion continues to apply. Lastly, in column (6), we include two dummy variables to capture whether the government is a minority or majority owner (with no government ownership, which represents the case for 94% of the firms in our data, as the omitted category). While there is weak evidence that when the government owns a minority stake in the firm, there is an increase in R&D expenditures by 8.7% (significant at the 10% level), all of the other effects (those of which we are primarily interested in) remain unchanged.

As our simple theoretic model based off Dasgupta and Stiglitz [1980] showed, R&D and market structure are determined simultaneously in equilibrium. Our reduced-form approach does not address these endogeneity concerns as covariates are likely to be correlated with the error term ε_{ijlt} in equation (7) due to strategic behavior, as our model presented earlier highlighted. Additionally, variables like the LI and number of employees are likely to be endogenous. A common approach to addressing endogeneity would to to adopt an IV regression model. To gauge the importance of endogeneity within a comparable framework, we instrument for market concentration, competition, and rivals' R&D.

The set of instruments that we use can be partitioned into two categories: those that are included and those that are excluded. The exogenous regressors used in the reduced-form model (firm's age, status as an exporter, variables capturing the number of products the firm manufactures, its' ownership structure, as well as the variables accounting for the number of (regional and national) universities in the same area, along with firm fixed effects) are considered to be included instruments. When it comes to excluded instruments, practitioners might use something like lagged versions of the independent variables (e.g., Arellano-Bond, which is not ideal given our short panel), or introduce new variables that can serve as instruments. For example, in our application, one might consider local/county tax rates or the distance to the nearest port. However, again because we have a short panel, taxes are time invariant and variables like distance are already being captured by firm fixed effects. As such, we create excluded instruments by constructing functions of the included instruments that are non-binary variables. Specifically, we square a firm's age, share of ownership, and the number of national and provincial universities. In addition, we interact these squares with all other exogenous binary regressors. These variables are

transformations of the covariates included in the OLS regression except for the main "problematic" variables (viz. rivals' R&D, LI, number of rivals). This follows from standard GMM practice where if a variable in an equation of interest is exogenous, functions of it can be used as a valid instrument. Constructing functions of the exogenous regressors to use as instruments has precedent; see, for example, Atkinson and Tsionas [2016] and Atkinson, Primont, and Tsionas [2018]. In our model we have endogeneity concerns surrounding seven regressors. We therefore require at least seven excluded instruments to identify the model. The variables created by squaring and interacting with each other easily exceeds seven (the number of included endogenous variables) leading to a model that is overidentified.

We estimate an IV regression that corresponds to model (6) using GMM and present the coefficient results in column (7) of Table 2.[14] Additionally, for easier comparison to the OLS estimates, Figure 4A depicts the estimated polynomials of the LI on the log of R&D expenditures under both approaches. The polynomial again suggests a nonlinear relationship between the measure of market concentration and R&D expenditures, though many of the coefficients themselves lose significance in the IV framework. Regardless, our primary point is that accounting for endogeneity seems to suggest very different relationships between market structure and R&D expenditures.

Of course, our reduced-form approach is in and of itself somewhat misspecified given the potential for fitted values to be negative. A strategy to deal with this in a reduced-form spirit would be to consider censored regression models such as this Tobit-like model:

$$\log\left(\mathrm{R\&D}_{ijlt}\right) = \begin{cases} 0, \text{ if } R\&D_{ijlt} \text{ is not available,} \\ f\left(\mathrm{LI}_{ijt}\right) + g\left(\mathrm{competition}_{ijlt}\right) + \alpha \log\left(\mathrm{rivals'\ R\&D}_{ijlt}\right) + \beta X_{it} + \gamma W_{lt} + \delta_i + \tau_t + \varepsilon_{ijlt}, \text{otherwise.} \end{cases} \tag{8}$$

Alternatively, a reduced-form approach might employ something like the Poisson pseudo maximum likelihood (PPML) approach suggested by Santos Silva and Tenreyro [2006] who apply the estimator to trade flow data. We opted to use firm fixed effects

[14] One concern readers might have relates to the validity of the instruments we employ. Hence, we also estimate our regressions using traditional IV regression, which provides *F*-statistics for the first stage regression. These statistics are substantially higher than the the Stock-Yogo *F*-test critical value of 20.74 (for a 5% level of bias relative to OLS). Hence, we have sufficient statistical evidence to reject the null hypothesis that the instruments are weak. We provide these results in Table 3.

in these specifications, which would be computationally difficult when considering PPML or a censored specification. Regardless, we recognize the censoring issue and address it in our Bayesian strategy.[15]

In addition, R&D expenditures of rival firms is, clearly, a noisy variable perhaps measured with error given we do not observe all rival firms (only firms with annual sales of at least five million RMB). The measurement error problem is attenuated by the fact that the specification we (and the literature) consider is nonlinear in the variables. Our main contribution is to adopt an alternative estimation strategy based on recent advancements in econometrics aimed at dealing with measurement error and endogenous regressors. We apply this alternative and estimate a corresponding model in the next section which can be compared with the baseline regression results from Table 2.

5. Bayesian Estimation Strategy

In this section, we adopt a framework which integrates a strategy for dealing with measurement error concerns (ELVIS) with a Bayesian version of GMM (BETEL) to deal with endogeneity issues. To take account of measurement error in nonlinear variables (like the polynomials involving the LI and rivals' R&D as well as the number of rivals) which also have endogeneity issues, we adapt two strategies proposed by Schennach [2005, 2014]. ELVIS is a general method to convert a model defined by moment conditions that involve both observed and unobserved variables into equivalent moment conditions that involve only observable variables allowing us to address measurement error concerns. BETEL uses these moment conditions and establishes the correct Bayesian posterior that should be used along with these moment conditions using instruments.

As such, our contribution is twofold, in the sense that we (i) deal with measurement error in critical variables of the model,[16] and (ii) we deal with endogeneity of rivals' R&D. We use a coherent and unified framework organized by combining ELVIS with BETEL in a general formulation which, in turn, can be

[15]All our Bayesian techniques go through using the likelihood function of (8) and flat priors for the parameters. We take account of measurent error in the non-zero observations but we assume the left censoring values at zero are correct. This assumption is not restrictive as zero means the firm did not disclose information which, in effect, means that these variables are truly zero.

[16]For example, measurement error may exist in the raw data or computation of the number of unique industries, the number of unique firms, the R&D data itself, operating profit, financial cost, sales, LI, number of rivals or rivals' R&D, the number of employees, or age of the firm.

estimated using fast MCMC-based methods. Remember, we also require addressing the challenge that our dependent variable is censored. To understand how our integrated framework tackles each of these issues, we continue by discussing each of these in more detail to provide an overview of our empirical strategy. We expand on this summary in Appendix B where we provide technical characterizations of ELVIS and BETEL.

Step 1:

The model in (8) is a censored regression model of the form

$$y_i^* = x'_i\beta + \varepsilon_i, \varepsilon_i|x_i \sim \text{i. i. d}\,\mathcal{N}(0,\sigma^2), i = 1, \dots, N, \quad (9)$$

where, in the interest of simplicity, we have omitted multiple subscripts and nonlinearities. In this expression, $y_i = \max(0, y_i^*)$, x_i is a vector of regressors (for which we assume, momentarily that they are not measured with error), β is a parameter vector, and ε_i is an error term, normally distributed with zero mean and standard deviation σ. Here, y_i is observed and, for simplicity but without loss of generality, $y_i > 0$ (i.e. log R&D is positive when R&D is observed). Suppose, again without loss of generality, that $y^* = [y_1, \dots, y_n, y_{n+1}^*, \dots, y_N^*]'$ so that the first $n < N$ observations contain data for which R&D is available.

If we knew the complete data in y^*, inference on β and σ would be straightforward, as we have a standard normal linear regression model. The unobserved data $y_{n+1}^*, \dots, y_N^*$ are also generated by (9) but they are subject to the restriction

$$y_i^*|x_i, \beta, \sigma \sim \mathcal{N}(x'_i\beta, \sigma^2), \text{ s. t. } y_i^* < 0, i = n+1, \dots, N. \quad (10)$$

Given x_i, β and σ, the missing values y_i^* can be generated from a normal distribution $\mathcal{N}(x'_i\beta, \sigma^2)$, subject to the restriction that the outcome is negative. This characterization corresponds to the Gibbs sampler data augmentation scheme (see, for example, Chib [1992]). The Gibbs sampler draws, successively from the so-called full conditional posterior distribution of $\beta|y^*$, $\sigma|y^*$ and, of course, $y_i^*|\beta, \sigma$. This generates a MCMC sample which converges to the posterior of the model. So, the crucial step of censoring can be easily dealt with using Gibbs sampling with data augmentation.

Step 2:

As some regressors may be measured with error we write

$$y_i^* = x_i^{*'}\beta_1 + z'_i\beta_2 + \varepsilon_i, \varepsilon_i|z_i \sim \text{i. i. d}\,\mathcal{N}(0,\sigma^2), i = 1,\dots,N, \tag{11}$$

where z_i is a vector of exogenous regressors. Specifically, in our case the z_is are firm's age, exporter status, indicators for the number of products a firm manufactures, share of the firm owned by the government, and the number of universities. The vector of variables x_i^* is measured with error, say

$$x_i = x_i^* + U_i, i = 1,\dots,N,$$

where x_i^* is the actual value, x_i is the observed value, and U_i is an error. Let $w_i = [y'_i, x'_i]'$.

Moment conditions have the general form:

$$\mathbb{E}[g(U,y,\theta)] = 0 \tag{12}$$

where g is a d_g-dimensional vector of nonlinear measurable functions depending on the parameter $\theta \in \Theta \subseteq \Re^k$. In this general notation, the unobserved random vector U takes values in $\mathcal{U} \subseteq \Re^{d_u}$ and the observed random vector takes value in $\mathcal{Y} \subseteq \Re^m$. Intuitively, the unobservables U can be eliminated from the moment condition by averaging the function $g(U,Y,\theta)$ over U. ELVIS is a method to convert a model defined by moment conditions that involve both observed and unobserved variables into equivalent moment conditions that involve only observable variables by "integrating out" the unobservables.

In our case, the function $g(\cdot)$ from (12) would be

$$g(U,y,\theta) = \begin{bmatrix} x_i - U_i \\ y_i^* - U'_i\beta_1 - z'_i\beta_2 \\ U_i \otimes (x_i - U_i) \\ U_i \otimes (y_i^* - U'_i\beta_1 - z'_i\beta_2) \\ (x_i - U_i) \otimes (y_i^* - U'_i\beta_1 - z'_i\beta_2) \\ z_i \otimes (y_i^* - U'_i\beta_1 - z'_i\beta_2) \\ z_i \otimes (x_i - U_i) \end{bmatrix}, \tag{13}$$

see Schennach [2014]. The only difference relative to Schennach [2014] is that (i) we have the additional, plausibly exogenous variables z_i that can act as instruments, and (ii) these instruments are orthogonal to measurement errors U_i in x_i.

The key is that ELVIS delivers a new, equivalent set of moment conditions based on observables alone, say

$$\mathbb{E}\tilde{g}(y,\theta,\gamma) = 0,$$

that do not contain the U_is although they may contain new (nuisance) parameters denoted by γ. As Schennach [2014] mentioned: "Although the unobservables have been 'integrated out' from the original moment conditions, the resulting averaged moment conditions are formally equivalent to the original moment conditions, in the sense that the values of parameters that solve the averaged moment conditions are the same as the values that solve the original moment conditions." In fact, Schennach [2014] describes the measurement error case in her Example 1.5. The method requires specifying a (dominating conditional) measure for the distribution of the unobservables given the observables. The exact choice has no effect on the results as long as it satisfies basic properties stated in Definition 2.2 and Proposition 2.1 of Schennach [2014]. Using a measure in the exponential family satisfies the so-called "least favorable" property which roughly says that the dominating measure is not simpler compared to the actual distribution of the unobservables given the observed variables. This averaging of the moment functions over unobservables should employ a least-favorable distribution which is obtained through an entropy maximization procedure which can be carried out through simulations, avoiding the need to make distributional assumptions about U_i. Technical details of this ELVIS procedure are explained in Appendix B.1.

Step 3:

Given the new moment conditions $\mathbb{E}\tilde{g}(y,\theta,\gamma) = 0$, one can use a number of different empirical strategies. One approach to estimating such models would be to employ GMM as developed by Hansen [1982]. However, a GMM-based approach may behave erratically in finite samples (see, for example, Li and Zhang [2009] and Barbosa et al. [2022]). Alternatively, estimating the parameters of interest and certain other nuisance parameters can be implemented through a variety of standard techniques like empirical likelihood (EL) or exponentially tilted empirical likelihood (ETEL), which provide more efficient estimates with reduced small-sample bias (Newey and Smith [2004], Schennach [2007]); see also Kitamura [2001], Kitamura et al. [2012], and Canay [2010]. However, if prior information is available and we want to condition on the observed data, then a Bayesian approach has appeal.

BETEL is essentially a nonparametric Bayesian version of GMM and it is, conceptually, closely related to EL and similar techniques where a sister likelihood

function allows for inference but does not require distributional assumptions. EL approaches reweight the sample so as to satisfy moment conditions exactly. BETEL employs a noninformative prior on the space of distributions, then uses moment conditions to establish the correct Bayesian posterior that should be used along with these moment conditions. This Bayesian procedure admits an EL-type representation where probability weights are selected via exponential tilting. Its implementation is computationally convenient and involves use of MCMC techniques which provide access to the posterior implied by moment conditions, as suitable instruments are available in our case. Thus, as Schennach [2005] noted, BETEL has the advantage of adapting to the "shape" of the data, unlike approaches based on the assumption of asymptotic normality. Technical details of this BETEL procedure are explained in Appendix B.2.

5.3. Bayesian Empirical Results

We implement our estimation strategy in Fortran 77 using the netlib and GNU Scientific Library subroutines. Our overall hybrid algorithm uses 150,000 iterations in all of our MCMC runs with a burn-in phase of 50,000 to mitigate the possible impact of starting values which are taken from least-squares dummy variables (LSDV) estimation. Standard Metropolis–Hastings algorithms may be quite inefficient computationally and the approach suggested by Schennach [2014] involves considerable tuning.[17] Therefore, we prefer to use the Girolami and Calderhead [2011] method described in Appendix B.3. The technique is reliable, requires almost no tuning and the MCMC draws it provides have considerable less autocorrelation compared to other MCMC algorithms.

We estimate models of the form provided in equation (7) under the specifications in Table 2, we provide analogous results in Table 4. Comparing the coefficient estimates in Tables 2 and 4 shows stark differences. The effect of rivals' average R&D, the size of the firm (as proxied by employment), exporter status, and the effect from being a multi-product firm are all reasonably similar. However, and importantly, the factors we've argued suffer from endogeneity have changed substantially. As an example, consider the polynomial representation of the LI measure.

[17]This is due to very high autocorrelations produced by the algorithm in most of its versions—for example, independence or random walk samplers.

Again, because polynomial coefficients are not very convenient to interpret, we depict the fitted polynomials under the Bayesian estimates in Figure 4, Panel B. The Bayesian estimate suggests the more market power a firm has, the more it invests in R&D—a trend supportive of Schumpeter's hypothesis, the theoretical model we presented earlier, and contrasting the nonlinear relationships suggested by our reduced form results (as well as those of researchers in this literature) presented in Figure 4, Panel A.

To support our creation of instruments using transformations of the included exogenous variables and in the spirit of computing a *J*-statistic, we consider the Hansen-Sargan criterion within our hybrid (BETEL-based rather than traditional GMM) approach for each model involving these instruments.[18] To be clear, *J*-statistics cannot be directly computed within our hybrid approach. As such, consider the following:

Suppose we have the moment conditions

$$\underset{(\mathrm{M}\times 1)}{\mathbf{g}}(\theta) = (\mathrm{nT})^{-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{g}_{\mathrm{it}}(\boldsymbol{\theta}), 1 \leq \mathrm{t} \leq \mathrm{T},$$

and the GMM criterion

$$\mathrm{J}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta})'\mathbb{W}(\boldsymbol{\theta})\mathbf{g}(\boldsymbol{\theta}),$$

where $\mathbb{W}(\boldsymbol{\theta})$ is the optimal weight GMM matrix as used in the Continuously Updated Estimator of GMM. For MCMC draw $\boldsymbol{\theta}^{(\mathrm{s})}$ $(1 \leq \mathrm{s} \leq \mathrm{S})$, we can compute the criterion $\mathrm{J}(\boldsymbol{\theta}^{(\mathrm{s})})$ $((1 \leq \mathrm{s} \leq \mathrm{S})$. Asymptotically in T, $\mathrm{J}(\boldsymbol{\theta})$ follows a chi-squared distribution with degrees of freedom $\mathrm{df} = \mathrm{M} - \mathrm{d}_{\theta}$, where M is the number of moment conditions and d_{θ} the number of unknown parameters. For the moment conditions to hold it is necessary that the posterior distribution of $\mathrm{J}(\boldsymbol{\theta})$, which can be obtained from the posterior mean of $\mathrm{J}(\boldsymbol{\theta}^{(\mathrm{s})})$ $((1 \leq \mathrm{s} \leq \mathrm{S})$ as, say $\mathrm{p}(\mathrm{J}(\boldsymbol{\theta})|\mathrm{D})$, where D is the entire data. Therefore, we need to test that $\mathrm{H}: \mathrm{J}(\boldsymbol{\theta}) = 0$. In turn, we compute p-values for the Hansen-Sargan criterion for each model we consider (these p-values correspond to the asymptotic chi-squared distribution of the test statistic). We fail to reject the null hypothesis that the overidentifying restrictions are valid at the standard (5%) level for every model.

[18]We include the same instruments employed in the GMM approach discussed earlier (results in column (7) of Table 2) as well as interactions of those instruments with firm dummies. Convergence diagnostics are available on request.

6. The Importance of Measurement Error and Endogeneity

While our empirical strategy sought to comprehensively address econometric challenges, it would be helpful to know the importance of measurement error and endogeneity on their own. In this section we do three things: we first address these issues individually by ignoring one step of our approach which shows that measurement error is most critical in our data; we then consider Monte Carlo simulations which deepen our understanding of these issues; lastly, we consider some robustness exercises.

6.1. Ignoring Measurement Error or Endogeneity

Our hybrid approach sought to address multiple issues at the same time. It's likely that both measurement error and endogeneity are at play, and perhaps compound each other. Regardless, to guage the relative importance of these issues and to help understand whether differences in our estimates relative to the reduced form results stem from measurement error or endogeneity, we consider ignoring one of these issues by shutting down one element of our hybrid approach. In Table 5, we present our baseline regression results from our OLS and GMM estimates in columns (1) and (2), respectively. In column (3), we ignore measurement error and address only endogeneity by adopting a BETEL approach that does not use the ELVIS-modified moment conditions. In column (4), we present estimates from using only ELVIS but ignoring endogeneity. In column (5), we present our full hybrid estimates. We have also included the polynomial fit from the BETEL-only and ELVIS-only models in Panel B of Figure 4, alongside the hybrid approach. In both the figure and in looking at the point estimates with respect to the LI-related terms, BETEL appears closer to our hybrid approach. However, the Bayes factors in the bottom of Table 5 offer relatively more support for ELVIS than BETEL (of course the full hybrid model has by far the most support) when considering the overall fit of the model. To investigate this, we consider some Monte Carlo and robustness exercises.

6.2. Monte Carlo Simulations

To help readers understand the important complications characterizing the environment

we study, we complement our estimates with Monte Carlo simulations motivated by our empirical specifications. We draw observations from our data to construct newly simulated datasets that we then estimate by imposing (or not imposing) various moment restrictions when using our hybrid estimation strategy. Conceptually, this is attractive as our full-blown estimation strategy essentially nests OLS which obtains when certain moment conditions are not imposed; likewise, we can impose a subset of the moment conditions to partially address issues like endogeneity or measurement error which helps shed light on the important complications that OLS neglects.

We generate data for a number of firms (n) and years (T) and we take the X matrix from the data by sampling data for n randomly selected firms. We assume competition is not contaminated with noise and we generate a contaminated version by adding a normal random variable with zero mean and standard deviation equal to its observed value (16.78 is the standard deviation on the number of rivals variable from Table 1). We have firm and year effects in all cases and we assume that the functional form of $f(\cdot)$ and $g(\cdot)$ in (7) is unknown but can be approximated using $h(x,z) = \sum_{i=1}^{p}\sum_{j=1}^{q} b_{ij}\, x^i z^j$, where x is LI, z is log number of rivals, p, q are the orders of the polynomial, and b_{ij} are unknown coefficients. In fact, we have $h(x,z) = f(x) + g(z)$ where $f(\cdot)$ and $g(\cdot)$ are second-order polynomials whose coefficients, like the coefficients of all other covariates, we take from column (6) of Table 4. In our Monte Carlo experiment the true polynomial orders (p,q) are unknown and we assume that $p,q \in \{1,\dots,5\}$. In turn, we generate the ys using the coefficient estimates from the Bayesian results (again, column (6) of Table 4), and adding a normal error with zero mean and standard deviation equal to the standard deviation of the Bayesian residuals.

To minimize computational costs, we use 60,000 MCMC iterations and we omit the first 10,000 to mitigate start up effects, if any. We are interested in whether we can find the truth about the functional forms $f(\cdot)$ and $g(\cdot)$.

It turns out that ELVIS works even when measurement error is substantial which is reassuring as this is one of the potential problems present in our data set. When the true model ($p = q = 2$) is not selected, the models most often selected are "over-fitted models" where p and/or q exceeds 2. This tendency disappears as sample size increases. We present summary results for various sample sizes concerning the number of firms and years in the simulated data in Table 6. Specifically, these results show the average root mean-squared error (ARMSE) and average absolute error (AAE) both fall

and the true model is selected more often, as the sample size increases.[19] In the same vein, Bayes factors (BF) increasingly select the correct model as sample size increases—either by increasing n or T.

To compare these results with OLS, as noted, we can apply ELVIS when we drop moment conditions that correspond to accounting for measurement error and/or endogeneity and then contrast ARMSE and AAE measures to the Bayesian results and recompute Bayes factors, essentially, allowing for model comparisons. The OLS results are reported in Table 7.[20] To compute these results, we use BETEL but drop the moment conditions corresponding to accounting for measurement error and endogeneity. The Bayes factors are quite large, thus favoring the ELVIS version with explicit account of measurement error and endogeneity, while the biases of ignoring them are quite substantial relative to the true values.

To make sense of these numbers consider the simple model $y = \beta x^* + u$ where $x = x^* + v$ and we only observe x instead of x^*. Suppose $x \perp u$. It can be shown easily that $\text{plim}\frac{\beta_{OLS}-\beta}{\beta} = -\frac{\sigma_v^2}{\sigma_x^2}$. When, as in our case, $\sigma_v = \sigma_x$ we would expect a downward bias roughly equal to 100%. As ELVIS is a highly nonlinear technique this cannot hold exactly but we see it is roughly correct, but at least helpful in giving context. Similarly, in Table 8 and Table 9, we shut down (drop/ignore) only the measurement error or endogeneity moment condition, respectively, and replicate the estimation process. While correcting for one of these concerns alone improves estimates, important biases remain if both issues are not addressed.

6.3. Robustness checks

Other researchers have noted that a firm's age or exporting status are also affected by a firm's productivity; see, for example, Olley and Pakes [1996] and De Loecker and Warzynski [2012]. Given we do not control for firm productivity directly (it is unobserved and thus omitted), one may be concerned about using age and exporting status as (exogenous) included instruments.[21] Therefore, we reconsider our empirical specification without age or exporting status included (they then become exluded

[19]In all of our Monte Carlo results, for parameters $\beta_1, \dots, \beta_k$ (denoting their true values) and posterior mean estimates $\hat{\beta}_1, \dots, \hat{\beta}_k$, $ARMSE = \sqrt{k^{-1}\sum_{j=1}^{k}(\hat{\beta}_j - \beta_j)^2}$ while $AAE = k^{-1}\sum_{j=1}^{k}|\hat{\beta}_j - \beta_j|$.

[20]Corresponding point estimates for the Monte Carlo simulations are available upon request.

[21] Note that a firm's markup (LI_{ijt}), by construction, is a function of firm costs which depend on firm-level productivity. Thus, firm productivity is accounted for indirectly in our approach.

instruments). In Table 10, we present estimation results for these models under the various empirical strategies we have considered. Qualitatively, the results are the same to the corresponding OLS, IV-GMM, and hybrid approach presented earlier, in Tables 2 and 4. Thus, we are not concerned about the validity of our proposed instruments.

A specification test that we can consider is to investigate whether f and g are separable in equation (7) instead of a more general model, say $h(LI_{ijt}, \text{competition}_{ijt})$. One way to do this, is assume that h can be approximated by using neural networks and then check the Bayes factors. The other alternative is to maintain the null hypothesis as the one in which f and g are separable polynomials of degree two, and use the test formalized by Horowitz [2006].

To understand this test, suppose $g(x)$ is a general non-parametric functional form and $G(x,\theta)$ is a parametric model with $\theta \in \Theta$. The null hypothesis of the test is $H_0: g(x) = G(x,\theta)$, for some $\theta \in \Theta$ and almost all x, and the alternative states that there does not exist a $\theta \in \Theta$ such that $g(x) = G(x,\theta)$ for almost every x. The model is $Y = g(x)$ and we assume that identification is achieved by $\mathbb{E}[Y - g(X)|W] = 0$, where W is an instrument for X. Therefore, we can rewrite the model as: $Y = g(X,Z) + U$, where $\mathbb{E}(U|Z,W) = 0$. We assume X, W are supported in $[0,1]^p$ and Z is supported in $[0,1]^r$ $(r \geq 0)$. Here, X and Z are endogenous and exogenous explanatory variables, respectively. Then, we need to test $H_0: g(x,z) = G(x,z,\theta)$, for known function G, some $\theta \in \Theta$ and almost every x, z. Suppose f_{XZW} denotes the joint density of random variables X, Z, W, f_z is the marginal density of Z and ν is some twice continuously differentiable function in $[0,1]^{p+r}$. We define a twice continuously differentiable operator such that

$$T_z \nu(x,z) = \int t_z(\xi, x)\nu(\xi, z)d\xi,$$

where $t_z(x_1, x_2) = \int f_{XZW}(x_1, z, w) f_{XZW}(x_2, z, w)dw$. Then H_0 is equivalent to

$$S(x,z) \equiv T_z[g(\cdot,\cdot) - G(\cdot,\cdot,\theta)](x,z) = 0,$$

for some $\theta \in \Theta$ and almost every x, z. Suppose $l(z_1, z_2)$ is the kernel of an operator

L in $[0,1]^r$ if $r > 0$, defined by $Lv(z) = \int l(\zeta, v)v(\zeta)d\zeta$. If $r = 0$ then L is the identity operator. The purpose of the operator is to carry out a smoothing step because when $r > 0$, the rate of convergence of the sample analogue of $\int S(x,z)^2 dxdz$ would be slower than $N^{-1/2}$ where N is the sample size. Under H_0,the sample analogue of $\int S(x,z)^2 dxdz$ can be constructed based on the observation that

$$T[g - G(\cdot,\cdot,\theta)](x,z) = E\{(Y - G(X,Z,\theta)\}f_{XW}(x,z,W)l(Z,z).$$

In turn, if we define $V_i = (X_i, Z_i, W_i)$, and $f_{XZW}^{(-i)}(v)$ denotes a leave-observation-i-out kernel estimator of f_{XZW} , and κ is a kernel in $\mathbb{R}^{2p+r}$, viz.

$$f_{XZW}^{(-i)}(v) = \frac{1}{nh^{2p+r}} \sum_{j=1, j\neq i}^{N} \kappa\left(\frac{v - V_i}{h}\right), \tag{14}$$

for bandwidth parameter h. Then we can construct the sample analogue as

$$\hat{S}(x,z) = n^{-1/2} \sum_{i=1}^{N} \left[Y_i - G(X_i, Z_i, \hat{\theta})\right] f_{XZW}^{(-i)}(x, Z_i, W_i) l(Z_i, z), \tag{15}$$

where $\hat{\theta}$ is a consistent estimator of θ. The test statistic is:

$$\tau = \int \hat{S}(x,z)^2 dxdz, \tag{16}$$

and H_0 is rejected when τ is large. To obtain asymptotic critical values, Horowitz [2006] shows that the asymptotic distribution is a mixture of χ^2 distributions, where the weights depend on the eigenvalues of a certain integral operator.

This test is important and has power against local alternatives whose distance from the null-hypothesis model is $O(N^{-1/2})$ where N is the sample size. We implement these tests by randomly sampling 10,000 firms at a time and we do this 1,000 times. We use 60,000 MCMC iterations and we omit the first 10,000 to mitigate possible start up effects. We use the same instruments as in our empirical application to keep things consistent and inform our empirical model specification. The results are

reported in Table 11. From the Bayes factors in favor of model (7) as well as the Horowitz [2006] test, it turns out that the specification in column (6) of Table 4 passes the diagnostic tests. We apply the tests to different sub-samples to investigate whether there is some form of misspecification.

From these results it does not seem that the specification in column (6) of Table 4 has substantial problems. A standard criticism of the Bayes factor is that it depends too much on the prior and it does not work well with improper priors. For this reason, we use the Hyvärinen [2005] score as implemented in Shao et al. [2019]. The results are reported in Table 12. The large values of the Hyvärinen [2005] score suggest that the model is, indeed, better compared to the other specifications.

7. Conclusion

In this paper, we adopt a hybrid approach to address censoring, measurement errors, and endogeneity jointly. We apply our econometric approach to examine the connections between firm size, market structure, and R&D efforts, where these relationships have been of interest to economists since (at least) Schumpeter [1942]. Researchers have long recognized that these relationships are difficult to disentangle because of endogeneity concerns. We motivate our setting by adapting a simple theoretical model based on Dasgupta and Stiglitz [1980] work where determinants of R&D investment decisions, market structure, and the number of firms active in a market are simultaneously decided. In addition, measurement error is likely a concern as there is a correlation between explanatory variables and the error term in econometric models. To these points, Aghion et al. [2014] noted "Moreover, clean and direct measures of innovation and competition are usually not available in field data, which can lead to the additional problem of measurement error."

We apply our econometric approach to Chinese manufacturing data that has been widely used in the literature. As we do not employ field data nor observe all firms in the population, our measures of rival firms' expenditures on R&D suffer from mismeasurement. Further, compounding these issues is the fact that the underlying relationships are likely nonlinear. Indeed, our OLS estimates suggest a nonlinear relationship between market structure (as measured by a LI) and the amount a firm spends on R&D. This finding is consistent with what other researchers have found using data from different time periods, countries, and industries. When we estimate the same model under our hybrid approach, we get a coherent and unified framework

organized around BETEL combined with ELVIS in a general formulation estimated using fast MCMC-based techniques. We find that there is a positive relationship between market concentration and R&D expenditures, suggesting that firms in more concentrated industries invest more in R&D. Our finding is consistent with Schumpeter's original hypothesis as well as with the simple theoretic model we use to highlight issues of simultaneity. We provide Monte Carlo evidence that supports our approach and emphasizes the presence of both measurement error and endogeneity in the data and helps explain why OLS is ill-suited for estimation in this setting.

The natural direction for future research would be to consider this hybrid Bayesian framework we have suggested for other types of problems or in other data sets. The differences in the estimates obtained under a traditional strategy and one that corrects for the issues we've highlighted suggests that concerns around simultaneity and measurement error are nontrivial, meaning previous findings may be worth revisiting. The R&D literature usually focuses on one of two things—efforts (often measured by expenditures as we have done) and output (often measured by patents or innovations). Our approach could be extended to outcome-based models and measures where endogeneity concerns remain, and variables are likely to be even more plagued by measurement error. More generally, we think this approach will be helpful in addressing environments with interdependencies amongst decision makers. For example, if players in a game maximize expected payoffs but there is an unobservable disturbance (seen by players but not the econometrician) that enters expected payoffs in nonlinear and even nonmonotone ways, our approach can be extended. We've considered that here with a focus on R&D, but other traditional examples include price-setting games, firm advertising decisions, product differentiation models, or tariff setting. While many researchers often use a strategy that requires lagged variables in addressing these concerns, we feel our approach is particularly attractive for short panel data with a large cross section of observations.

References

Philippe Aghion, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics*, 120(2):701–728, 2005.

Philippe Aghion, Stefan Bechtold, Lea Cassar, and Holger Herz. The causal effects of competition on innovation: Experimental evidence, Harvard University, Department of Economics, typescript, 2014.

Gautam Ahuja, Curba Morris Lampert, and Vivek Tandon. Moving beyond Schumpeter: Management research on the determinants of technological innovation. *Academy of Management Annals*, 2(1):1–98, 2008.

Scott E. Atkinson and Mike G. Tsionas. Directional distance functions: optimal endogenous directions. *Journal of Econometrics*. 190(2): 301-314, 2016.

Scott E. Atkinson, Daniel Primont, and Mike G. Tsionas. Statistical inference in efficient production with bad inputs and outputs using latent prices and optimal directions. *Journal of Econometrics*. 204(2): 131-146, 2018.

Chong-En Bai, Jiangyong Lu, and Zhigang Tao. The multitask theory of state enterprise reform: Empirical evidence from China. *American Economic Review*, 96(2):353–357, 2006.

Klenio Barbosa, Dakshina G. De Silva, Liyu Yang, and Hisayuki Yoshimoto. Auction Mechanisms and Treasury Revenues: Evidence from the Chinese Experiment. *American Economic Journal: Microeconomics*. 14(4):394-419, 2022.

Rene´ Belderbos, Marcelina Grabowska, Stijn Kelchtermans, Bart Leten, Jojo Jacob, and Massimo Riccaboni. Whither geographic proximity? Bypassing local R&D units in foreign university collaboration. *Journal of International Business Studies*, 2021. https://doi.org/10.1057/s41267-021-00413-6

Richard Blundell, Rachel Griffith, and John van Reenen. Market share, market value and innovation in a panel of British manufacturing firms. *Review of Economic Studies*, 66(3):529–554, 1999.

Hongbin Cai and Qiao Liu. Competition and corporate tax avoidance: Evidence from Chinese industrial firms. *The Economic Journal*, 119(537):764–795, 2009.

Ivan A. Canay. EL inference for partially identified models: Large deviations optimality and bootstrap validity. *Journal of Econometrics*, 156(2):408–425, 2010.

Siddharta Chib. Bayes inference in the Tobit censored regression model. *Journal of Econometrics* 51(1-2):79-99, 1992.

Wesley M. Cohen. Empirical studies of innovative activity. volume 1 of *Handbook of the Economics of Innovation and Technical Change*, pages 182–264. Blackwell, 1995.

Wesley M. Cohen. Fifty years of empirical studies of innovative activity and performance. In Bronwyn H. Hall and Nathan Rosenberg, editors, *Handbook of The Economics of Innovation*, Vol. 1, volume 1 of *Handbook of the Economics of Innovation*, pages 129–213. North-Holland, 2010.

Wesley M. Cohen and Richard C. Levin. *Empirical studies of innovation and market structure*. Volume 2 of Handbook of Industrial Organization, pages 1059–1107. Elsevier, 1989.

Robert A Connolly and Mark Hirschey. R&D, market structure, and profits: A value-based approach. *Review of Economics and Statistics*, 66(4):682–686, 1984.

J. Danielsson and J.-F. Richard. Accelerated Gaussian importance sampler with application to dynamic latent variable models. *Journal of Applied Econometrics*, 8(S1):S153–S173, 1993.

Partha Dasgupta and Joseph Stiglitz. Industrial structure and the nature of innovative activity. *The Economic Journal*, 90(358):266–93, 1980.

Jan De Loecker and Frederic Warzynski. Markups and Firm-Level Export Status. *American Economic Review*, 102(6):2437–2471, 2012.

Dakshina G. De Silva, Thimothy P. Hubbard, Anita. R. Schiller. Entry and exit patterns of 'toxic' firms. *American Journal of Agricultural Economics*. 98 (3), 881–909, 2016.

Dakshina G. De Silva, Robert P. McComb, Anita R. Schiller, and Aurelie Slechten. Firm Behavior and Pollution in Small Geographies. *European Economic Review*, 136, 2021 https://doi.org/10.1016/j.euroecorev.2021.103742

Paul Geroski. Innovation, technological opportunity, and market structure. *Oxford Economic Papers*, 42(3):586–602, 1990.

Richard Gilbert. Looking for Mr. Schumpeter: Where are we in the competition–innovation debate? *Innovation Policy and the Economy*, 6:159–215, 2006.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), 73(2): 123–214, 2011.

Gene M. Grossman and Alan Krueger. Economic Growth and the Environment. *The Quarterly Journal of Economics*, 110 (2): 353-377, 1995

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

Joel L. Horowitz. Testing a Parametric Model against a Nonparametric Alternative

with Identification through Instrumental Variables. *Econometrica*, 74, 521-538, 2006.

Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6: 695–709, 2005.

Yuichi Kitamura. Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, 69(6):1661–1672, 2001.

Yuichi Kitamura, Andres Santos, and Azeem M. Shaikh. On the asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica*, 80(1):413–423, 2012.

Richard Levin and Peter C. Reiss. Tests of a Schumpeterian model of r&d and market structure. In *R&D, Patents, and Productivity*, NBER Chapters, pages 175–208. National Bureau of Economic Research, Inc, 1984.

Richard C. Levin and Peter C. Reiss. Cost-reducing and demand-creating R&D with spillovers. *RAND Journal of Economics*, 19(4):538–556, 1988.

Richard C Levin, Wesley M Cohen, and David C Mowery. R&D appropriability, opportunity, and market structure: New evidence on some Schumpeterian hypotheses. *American Economic Review*, 75(2):20–24, 1985.

Jiatao Li, Haoyuan Ding, Yichuan Hu, and Guoguang Wan. Dealing with dynamic endogeneity in international business research. *Journal of International Business Studies* 52, 339–362, 2021.

Tong Li and Xiaoyong Zheng. Entry and Competition Effects in First-Price Auctions: Theory and Evidence from Procurement Auctions. *Review of Economic Studies* 76 (4): 1397–1429, 2009.

Jiangyong Lu and Zhigang Tao. Trends and determinants of China's industrial agglomeration. *Journal of Urban Economics*, 65(2):167–180, 2009.

Whitney K. Newey and Richard J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

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

Stuart S. Rosenthal and William C. Strange. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics*, 85(2):377–393, 2003.

J. M. C. Santos Silva and Silvana Tenreyro. The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658, 2006.

Susanne M. Schennach. Bayesian Exponentially Tilted Empirical Likelihood. *Biometrika*, 92(1): 31–46, 2005.

Susanne M. Schennach. Point estimation with exponentially tilted empirical likelihood.

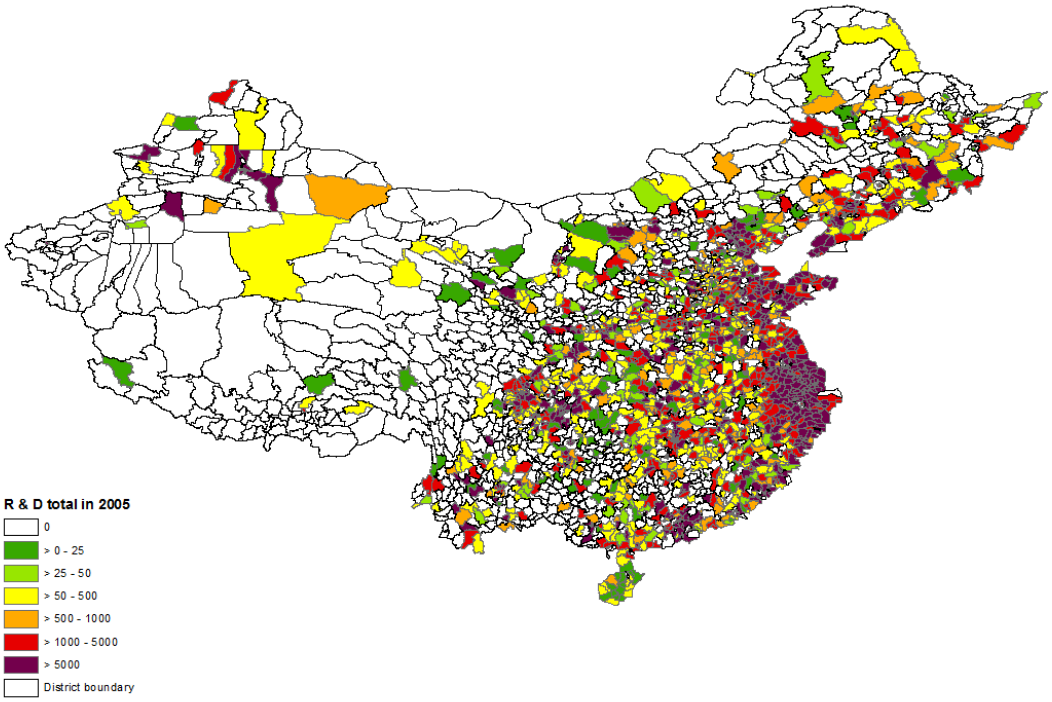
The Annals of Statistics, 35(2):634–672, 2007.

Susanne M. Schennach. Entropic latent variable integration via simulation. *Econometrica*, 82(1): 345–385, 2014.

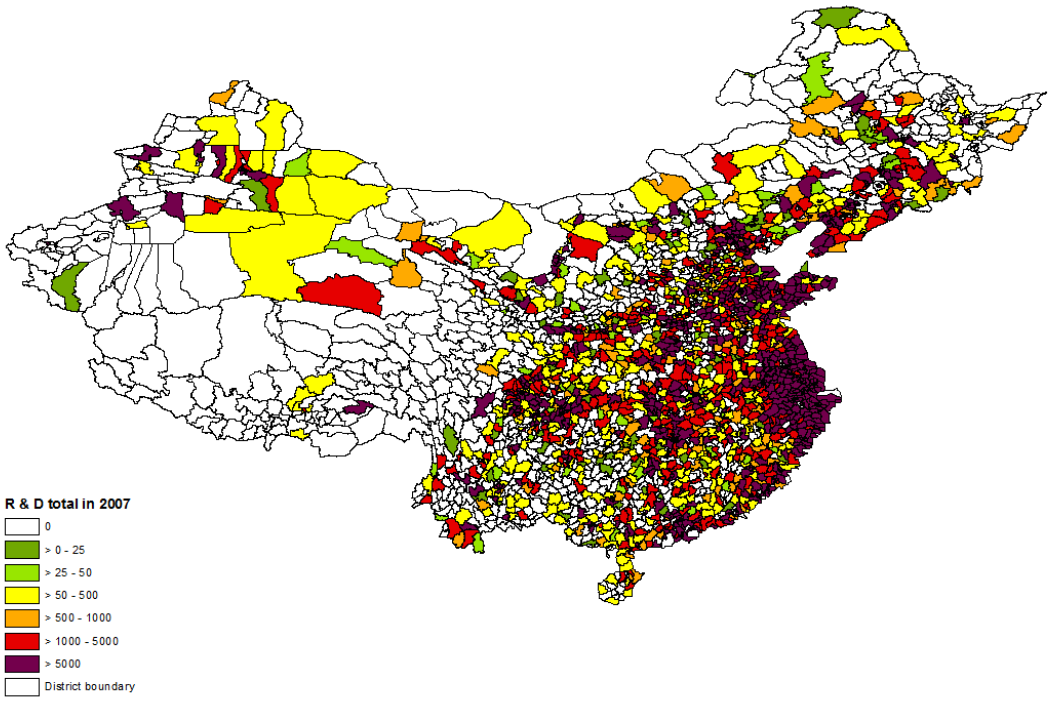
Frederic M. Scherer. Market structure and the employment of scientists and engineers. *American Economic Review*, 57(3):524–531, 1967.

Joseph A. Schumpeter. *Capitalism, Socialism, and Democracy*. Harper, New York, New York, 1942.

Robert W Wilson. The effect of technological environment and product rivalry on R&D effort and licensing of inventions. *Review of Economics and Statistics*, 59(2):171–178, 1977.



Panel A: 2005



Panel B: 2007

Figure 1: Total R&D Expenditures by District

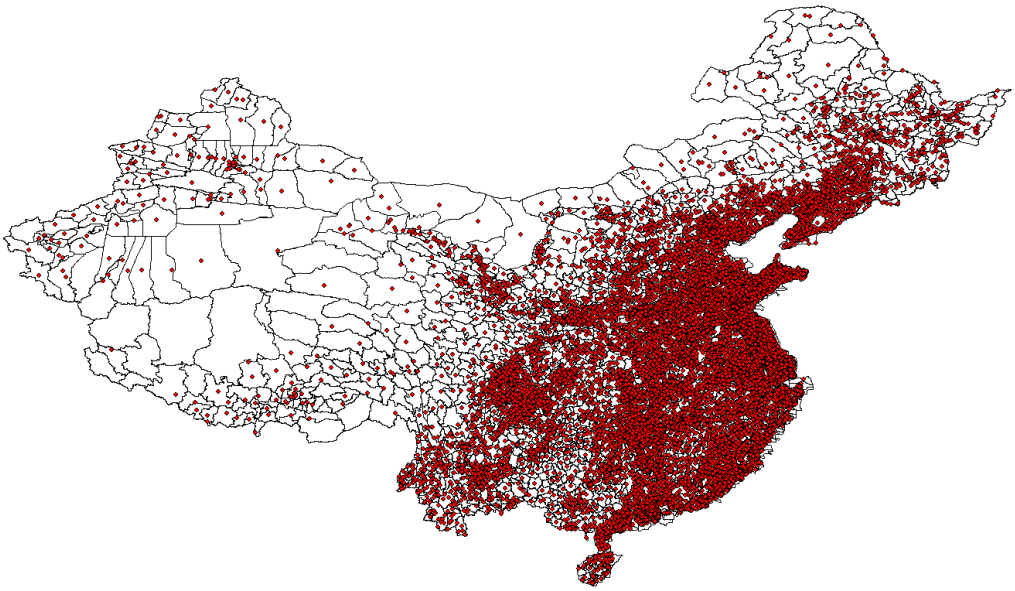


Figure 2. Manufacturing Firm Locations in Sample

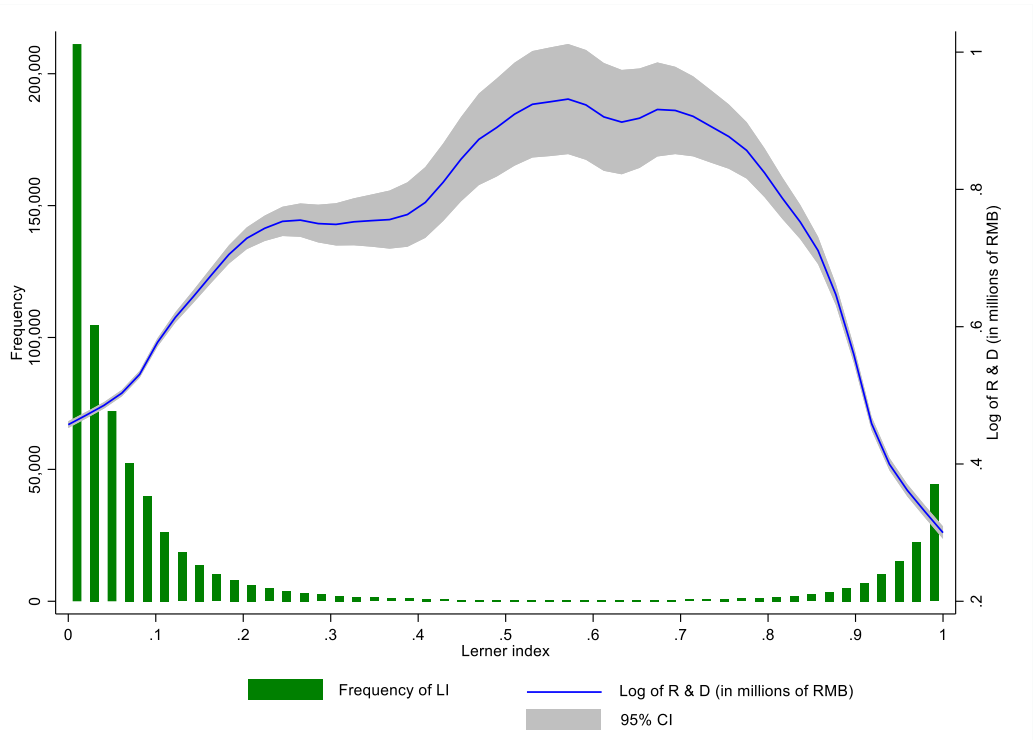
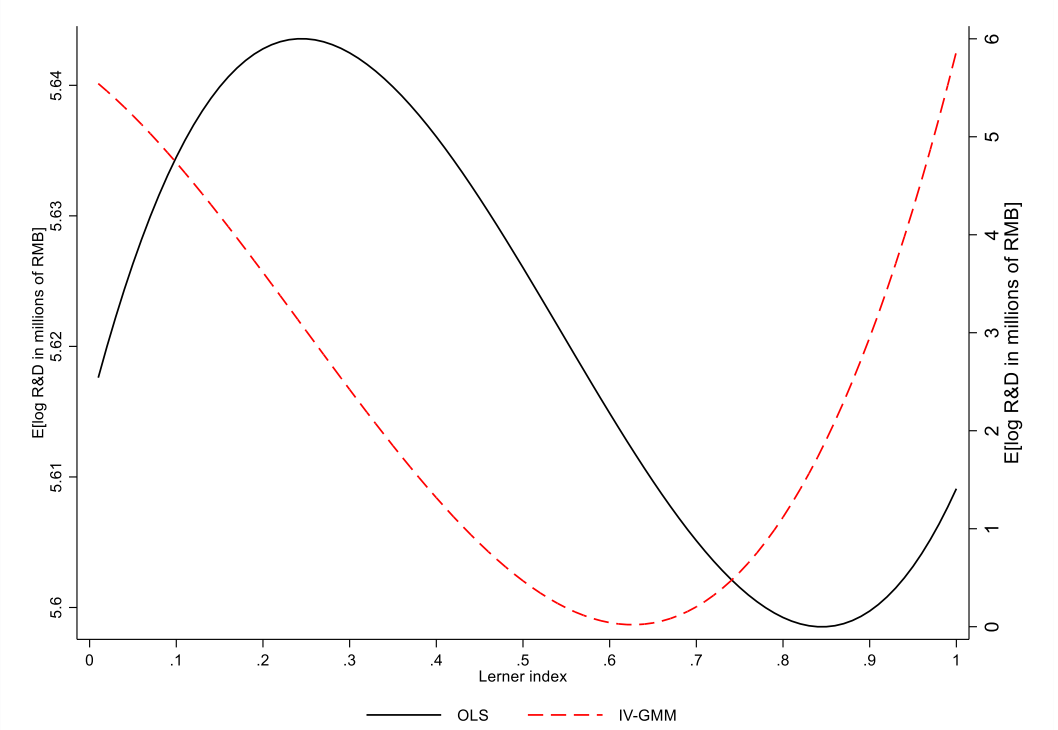
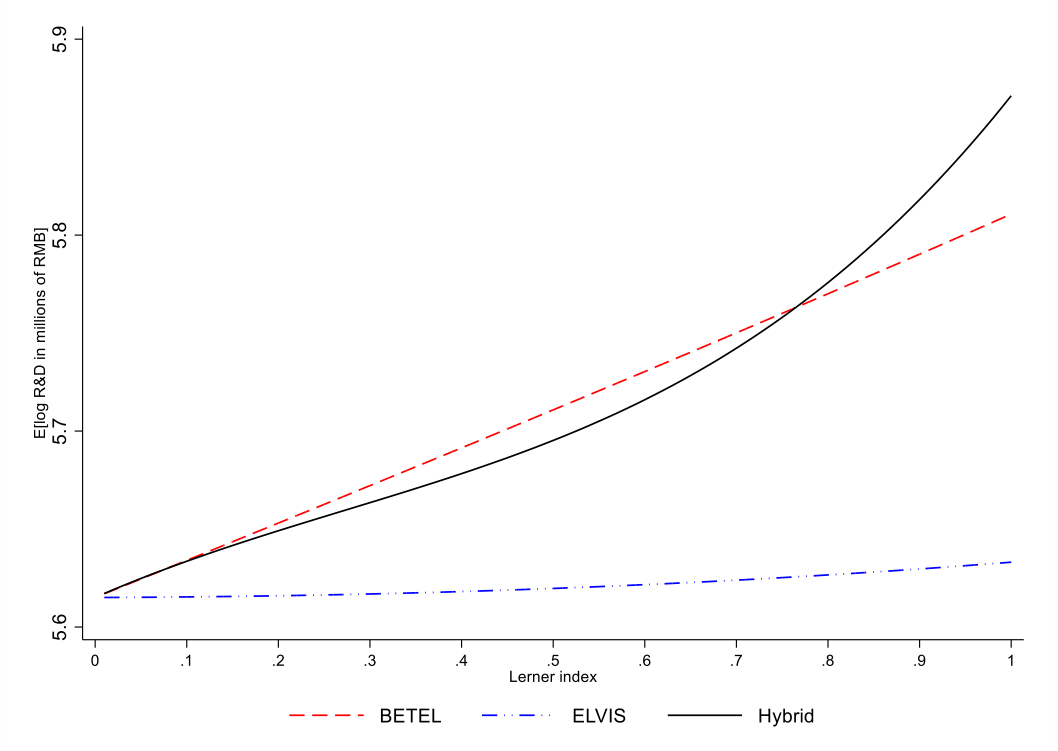


Figure 3. Histogram of LI Measure and Plot Against Log(R&D Expenditures)



Panel A: Polynomial Estimates from OLS and IV-GMM
Note that IV results are plotted in the secondary y-axis.



Panel B: Polynomial Estimates from Bayesian Approach

Figure 4: R&D and LI

Table 1: Summary statistics

Variables	Data	
	Sample	$\pi_{it} \geq 0$
Number of unique industries (at 4–digits)	525	525
Number of unique firms	324,463	319,609
R & D[a]	269.613	274.539
	(7,909.756)	(8,054.781)
Operating profit[a]	14,347.950	14,877.160
	(192,159.380)	(195,954.100)
Financial cost[a]	875.231	761.099
	(9,665.995)	(8,557.175)
Current sales[a]	82,856.520	83,678.310
	(616,128.400)	(625,723.400)
LI	0.202	0.210
	(0.335)	(0.340)
Number of rivals	5.795	5.830
	(16.780)	(16.778)
Rivals' R & D[a]	135.827	136.759
	(4,422.454)	(4,489.228)
Number of employees	201.020	200.240
	(634.795)	(639.677)
Age	8.235	8.174
	(9.165)	(9.090)
Exporter	0.238	0.238
	(0.426)	(0.426)
Producing two outputs	0.240	0.239
	(0.427)	(0.427)
Producing three or more outputs	0.092	0.091
	(0.289)	(0.288)
Share of government ownership	0.049	0.048
	(0.207)	(0.204)
Share of government ownership: = 0	0.940	0.941
	(0.238)	(0.235)
Share of government ownership: > 0 – 0.50	0.012	0.012
	(0.110)	(0.109)
Share of government ownership: > 0.50	0.048	0.047
	(0.214)	(0.211)
Number of national universities in the same zip code	0.017	0.0170
	(0.227)	(0.229)
Number of provincial universities in the same zip code	0.214	0.213
	(.730)	(0.730)

Standard deviations are in parentheses. [a]: In millions of renminbi (RMB).

Table 2: Regression Results for R&D Spending when profits ≥ 0

Variables	Log of R & D_{ijlt}						
	OLS						IV-GMM
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
LI_{ijt}	0.258		0.256	0.257	0.257	0.257	-7.035
	(0.127)		(0.127)	(0.127)	(0.127)	(0.127)	(28.292)
LI_{ijt}^2	-0.680		-0.674	-0.678	-0.677	-0.679	-20.325
	(0.385)		(0.385)	(0.385)	(0.385)	(0.385)	(119.505)
LI_{ijt}^3	0.416		0.412	0.414	0.414	0.415	27.613
	(0.277)		(0.277)	(0.278)	(0.278)	(0.278)	(96.126)
Log number of $rivals_{ijlt}$		-0.087	-0.087	-0.086	-0.086	-0.086	1.521
		(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(4.863)
Log number of $rivals_{ijlt}^2$		0.038	0.038	0.038	0.038	0.038	-1.209
		(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(3.085)
Log number of $rivals_{ijlt}^3$		-0.006	-0.006	-0.006	-0.006	-0.006	0.267
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.413)
Log of rivals' average R & D_{ijlt}	0.066	0.068	0.068	0.068	0.068	0.068	0.681
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.341)
Log number of $employees_{it}$	0.111	0.112	0.112	0.111	0.111	0.111	0.082
	(0.010)	(0.011)	(0.010)	(0.010)	(0.010)	(0.010)	(0.057)
Log of age_{it}	-0.005	-0.003	-0.003	-0.003	-0.003	-0.003	-0.044
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.033)
$Exporter_{it}$	0.099	0.099	0.099	0.099	0.099	0.099	0.024
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.031)
Producing two $outputs_{it}$				0.014	0.014	0.014	0.121
				(0.021)	(0.021)	(0.021)	(0.063)
Producing three or more $outputs_{it}$				-0.005	-0.005	-0.005	-0.038
				(0.032)	(0.032)	(0.032)	(0.042)
Share of government $ownership_{it}$					-0.034		
					(0.047)		
Share of government $ownership_{it}$: >0 – 0.50						0.087	0.095
						(0.049)	(0.051)
Share of government $ownership_{it}$: > 0.50						-0.032	-0.049
						(0.043)	(0.048)
Log number of national $universities_{lt}$				-0.204	-0.204	-0.204	-0.526
				(0.242)	(0.242)	(0.242)	(0.230)
Log number of provincial $universities_{lt}$				-0.074	-0.074	-0.074	-0.251
				(0.061)	(0.061)	(0.061)	(0.110)
Firm effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	710,639	710,639	710,639	710,639	710,639	710,639	710,639
R^2	0.794	0.794	0.794	0.794	0.794	0.794	
Hansen's *J*-statistic χ^2: *p*-value							0.1273

Robust standard errors clustered by industry are in parentheses.

Table 3: First-Stage IV Regression Statistics

Variable	**First stage *F*-statistic**
LI_{ijt}	266.92
LI_{ijt}^2	266.54
LI_{ijt}^3	263.52
Log number of $rivals_{ijlt}$	404.20
Log number of $rivals_{ijlt}^2$	448.80
Log number of $rivals_{ijlt}^3$	461.45
Log of rivals' average R & D_{ijlt}	126.07

F(10,710617)

Table 4: Bayesian Regression Results for Research and Development Spending when profits ≥ 0

Variables	Log of R & D_{ijlt}					
	(1)	(2)	(3)	(4)	(5)	(6)
LI_{ijt}	0.133		0.181	0.201	0.183	0.203
	(0.0071)		(0.0045)	(0.0350)	(0.0044)	(0.0230)
LI^2_{ijt}	-0.230		-0.189	-0.233	-0.184	-0.230
	(0.0070)		(0.0036)	(0.0071)	(0.0035)	(0.0071)
LI^3_{ijt}	0.255		0.317	0.289	0.314	0.287
	(0.0140)		(0.0210)	(0.0160)	(0.0250)	(0.0140)
Log number of $rivals_{ijlt}$		-0.014	-0.044	-0.032	-0.034	-0.034
		(0.0032)	(0.0019)	(0.0022)	(0.0016)	(0.0025)
Log number of $rivals^2_{ijlt}$		0.021	0.030	0.028	0.027	0.023
		(0.0044)	(0.0120)	(0.0035)	(0.0130)	(0.0032)
Log number of $rivals^3_{ijlt}$		-0.0033	-0.0028	-0.0030	-0.0024	-0.003
		(0.0010)	(0.0003)	(0.0005)	(0.0002)	(0.0040)
Log of rivals' average R & D_{ijlt}	0.044	0.052	0.045	0.033	0.043	0.034
	(0.0032)	(0.0130)	(0.017)	(0.0090)	(0.0120)	(0.0040)
Log number of $employees_{it}$	0.092	0.177	0.181	0.144	0.181	0.130
	(0.0130)	(0.0320)	(0.0170)	(0.0150)	(0.0120)	(0.0011)
Log of age_{it}	0.012	0.014	0.013	0.019	0.013	0.012
	(0.0032)	(0.0016)	(0.0021)	(0.0011)	(0.0020)	(0.0013)
$Exporter_{it}$	0.111	0.093	0.103	0.115	0.116	0.119
	(0.0034)	(0.0140)	(0.0045)	(0.0051)	(0.0032)	(0.0023)
Producing $two\ outputs_{it}$				0.021	0.023	0.020
				(0.0035)	(0.0035)	(0.0014)
Producing $three\ or\ more\ outputs_{it}$				0.083	0.071	0.073
				(0.0140)	(0.014)	(0.0070)
Share of government $ownership_{it}$					-0.0031	
					(0.0018)	
Share of government $ownership_{it}$: >0 – 0.50						0.020
						(0.0014)
Share of government $ownership_{it}$: > 0.50						-0.016
						(0.0021)
Log number of $national\ universities_{lt}$				0.045	0.030	0.035
				(0.0017)	(0.0022)	(0.0017)
Log number of $provincial\ universities_{lt}$				0.071	0.044	0.073
				(0.0130)	(0.016)	(0.0070)
Firm effects	Yes	Yes	Yes	Yes	Yes	Yes
Year effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	710,639	710,639	710,639	710,639	710,639	710,639
R^2	0.710	0.713	0.727	0.832	0.727	0.830
Bayes factor	1.000	1.616	4.183	41.630	1.202	52.833

Posterior standard deviations are in parentheses. Bayes factor has to been re-normalized to 1 using specification in Column 1.

Table 5: Regression Results for R&D Spending when profits ≥ 0

Variables	Log of R & D_{ijlt}				
	OLS	IV-GMM	BETEL	ELVIS	Hybrid
	(1)	(2)	(3)	(4)	(5)
LI_{ijt}	0.257	-7.035	0.189	0.0013	0.203
	(0.127)	(28.292)	(0.0012)	(0.0020)	(0.0230)
LI_{ijt}^2	-0.679	-20.325	0.0032	0.015	-0.230
	(0.385)	(119.505)	(0.0011)	(0.013)	(0.0071)
LI_{ijt}^3	0.415	27.613	0.0034	0.0017	0.287
	(0.278)	(96.126)	(0.0053)	(0.0032)	(0.0140)
Log number of $rivals_{ijlt}$	-0.086	1.521	0.003	-0.0027	-0.034
	(0.027)	(4.863)	(0.0021)	(0.0005)	(0.0025)
Log number of $rivals_{ijlt}^2$	0.038	-1.209	0.0045	0.0032	0.023
	(0.018)	(3.085)	(0.0012)	(0.0012)	(0.0032)
Log number of $rivals_{ijlt}^3$	-0.006	0.267	0.0013	0.0032	-0.003
	(0.003)	(0.413)	(0.0010)	(0.0012)	(0.0040)
Log of rivals' average R & D_{ijlt}	0.068	0.681	-0.0052	-0.0043	0.034
	(0.008)	(0.341)	(0.0010)	(0.0040)	(0.0040)
Log number of $employees_{it}$	0.111	0.082	0.144	0.053	0.130
	(0.010)	(0.057)	(0.0013)	(0.065)	(0.0011)
Log of age_{it}	-0.003	-0.044	0.0024	0.0015	0.012
	(0.014)	(0.033)	(0.0023)	(0.0017)	(0.0013)
$Exporter_{it}$	0.099	0.024	0.0044	-0.0040	0.119
	(0.019)	(0.031)	(0.0030)	(0.0012)	(0.0023)
Producing $two\ outputs_{it}$	0.014	0.121	0.0050	0.001	0.020
	(0.021)	(0.063)	(0.0034)	(0.0027)	(0.0014)
Producing $three\ or\ more\ outputs_{it}$	-0.005	-0.038	-0.0032	-0.003	0.073
	(0.032)	(0.042)	(0.0004)	(0.0082)	(0.0070)
Share of government $ownership_{it}$: >0 – 0.50	0.087	0.095	0.006	0.0020	0.020
	(0.049)	(0.051)	(0.0012)	(0.0032)	(0.0014)
Share of government $ownership_{it}$: > 0.50	-0.032	-0.049	-0.004	-0.004	-0.016
	(0.043)	(0.048)	(0.0001)	(0.0045)	(0.0021)
Log number of $national\ universities_{lt}$	-0.204	-0.526	0.0012	0.0045	0.035
	(0.242)	(0.230)	(0.0002)	(0.0024)	(0.0017)
Log number of $provincial\ universities_{lt}$	-0.074	-0.251	0.003	0.003	0.073
	(0.061)	(0.110)	(0.0001)	(0.0050)	(0.0070)
Firm effects	Yes	Yes	Yes	Yes	Yes
Year effects	Yes	Yes	Yes	Yes	Yes
Observations	710,639	710,639	710,639	710,639	710,639
R^2	0.794		0.533	0.599	0.830
Bayes factor			2.323	7.226	52.833
Hansen's *J*-statistic χ^2: *p*-value		0.1273			

Table 6: Monte Carlo Results: Bayesian

n and T	% true model selected by BF	ARMSE	AAE	AM
$n = 500,\ T = 10$	32	0.0322	0.0316	$(p, q) = (3,3)$
$n = 1000,\ T = 10$	75	0.0124	0.0120	$(p, q) = (3,2)$
$n = 10{,}000,\ T = 10$	93	0.0073	0.0072	$(p, q) = (2,3)$
$n = 500,\ T = 20$	82	0.0120	0.0121	$(p, q) = (3,3)$
$n = 1000,\ T = 20$	88	0.0093	0.0095	$(p, q) = (3,3)$
$n = 10{,}000,\ T = 20$	94	0.0054	0.0051	$(p, q) = (3,2)$

Notes: The true model is second-order polynomials for *f* and *g* in (7). ARMSE and AM are computed as noted in footnote 6 of the main text. AM stands for "alternative model" and corresponds to model selected most often when the true model is not selected.

Table 7: Monte Carlo Results: OLS

n and T	BF	OLS % ARMSE	% OLS AAE
$n = 500,\ T = 10$	3.32×10^{4}	130.5%	139.2%
$n = 1000,\ T = 10$	1.87×10^{5}	116.3%	122.3%
$n = 10{,}000,\ T = 10$	2.44×10^{7}	114.2%	114.3%
$n = 500,\ T = 20$	2.81×10^{4}	114.5%	114.0%
$n = 1000,\ T = 20$	4.50×10^{5}	114.2%	114.4%
$n = 10{,}000,\ T = 20$	1.17×10^{6}	114.2%	114.4%

Notes: The true model is second-order polynomials for *f* and *g* in (7). ARMSE and AM are computed as noted in footnote 6 of the main text. AM stands for "alternative model" and corresponds to model selected most often when the true model is not selected.

Table 8: Monte Carlo Results: Correction for Endogeneity but not Measurement Error

n and T	% true model selected by BF	ARMSE	AAE	AM
$n = 500,\ T = 10$	3.3	46.18%	45.12%	$(p, q) = (2,2)$
$n = 1000,\ T = 10$	3.5	48.21%	48.00%	$(p, q) = (2,3)$
$n = 10{,}000,\ T = 10$	3.7	47.33%	47.30%	$(p, q) = (2,2)$
$n = 500,\ T = 20$	4.1	47.23%	47.25%	$(p, q) = (1,2)$
$n = 1000,\ T = 20$	4.5	48.25%	48.10%	$(p, q) = (1,2)$
$n = 10{,}000,\ T = 20$	5.6	48.21%	48.15%	$(p, q) = (2,2)$

Notes: The true model is second-order polynomials for *f* and *g* in (7). ARMSE and AM are computed as noted in footnote 6 of the main text. AM stands for "alternative model" and corresponds to model selected most often when the true model is not selected.

Table 9: Monte Carlo Results: Correction for Measurement Error but not Endogeneity

n and T	% true model selected by BF	ARMSE	AAE	AM
$n = 500,\ T = 10$	37.29%	28.32%	28.30%	$(p, q) = (2,1)$
$n = 1000,\ T = 10$	39.40%	26.44%	26.40%	$(p, q) = (2,2)$
$n = 10{,}000,\ T = 10$	43.20%	22.10%	22.15%	$(p, q) = (2,1)$
$n = 500,\ T = 20$	32.30%	34.12%	34.19%	$(p, q) = (2,1)$
$n = 1000,\ T = 20$	41.20%	24.30%	24.20%	$(p, q) = (2,1)$
$n = 10{,}000,\ T = 20$	45.00%	20.44%	20.32%	$(p, q) = (2,1)$

Notes: The true model is second-order polynomials for *f* and *g* (7). ARMSE and AM are computed as noted in footnote 6 of the main text. AM stands for "alternative model" and corresponds to model selected most often when the true model is not selected.

Table 10: Regression Results for R&D Spending when profits ≥ 0, alternate specification

Variables	Log of R & D_{ijlt}		
	OLS	IV-GMM	Hybrid
	(1)	(2)	(3)
LI_{ijt}	0.248	5.375	0.213
	(0.127)	(34.619)	(0.0240)
LI_{ijt}^2	-0.664	-112.669	-0.234
	(0.385)	(134.794)	(0.0069)
LI_{ijt}^3	0.410	109.966	0.280
	(0.278)	(104.254)	(0.0150)
Log number of $rivals_{ijlt}$	-0.085	5.617	-0.032
	(0.027)	(5.493)	(0.0021)
Log number of $rivals_{ijlt}^2$	0.036	-3.916	0.025
	(0.018)	(3.364)	(0.0030)
Log number of $rivals_{ijlt}^3$	-0.006	0.634	-0.004
	(0.003)	(0.444)	(0.0020)
Log of rivals' average R & D_{ijlt}	0.068	0.376	0.030
	(0.008)	(0.430)	(0.0027)
Log number of $employees_{it}$	0.114	0.053	0.127
	(0.011)	(0.071)	(0.0010)
Producing $two\ outputs_{it}$	0.014	0.146	0.017
	(0.021)	(0.083)	(0.0012)
Producing $three\ or\ more\ outputs_{it}$	-0.005	-0.074	0.070
	(0.032)	(0.048)	(0.0081)
Share of government $ownership_{it}$: >0 – 0.50	0.087	0.130	0.017
	(0.049)	(0.062)	(0.0012)
Share of government $ownership_{it}$: > 0.50	-0.032	-0.006	-0.018
	(0.043)	(0.056)	(0.0017)
Log number of $national\ universities_{lt}$	-0.206	-0.600	0.032
	(0.242)	(0.311)	(0.0014)
Log number of $provincial\ universities_{lt}$	-0.074	-0.312	0.069
	(0.061)	(0.095)	(0.0062)
Firm effects	Yes	Yes	Yes
Year effects	Yes	Yes	Yes
Observations	710,639	710,639	710,639
R^2	0.794		0.800
Bayes factor			46.325
Hansen's *J*-statistic χ^2: *p*-value		0.1453	

Table 11: Robustness checks

Sample	BF in favor of model in (7)[1]	
10,000 randomly selected firms[3]	87.43 (44.55–151.3)[2]	0.225 (0.073–0.377)[2]
Entire data set	125.12	0.280
First half of data set	93.40	0.115
Second half of data set	144,32	0.192
Lower 10% of rivals	104.71	0.117
Upper 10% of rivals	120.40	0.166
Lower 25% of rivals	130.21	0.170
Upper 25% of rivals	118.16	0.116

Notes: [1]Against a general neural network formulation whose number of nodes is determined by the BF. [2] Median value and 99% confidence interval. For the Horowitz [2006] we test we use the same moment conditions as in our implementation of ELVIS. [3]We implement these tests by randomly sampling 10,000 firms at a time and we do this 1,000 times. We use 60,000 MCMC iterations and we omit the first 10,000 to mitigate possible start up effects.

Table 12: Robustness checks (H score)

Sample	Hyvärinen score in favor of model in (7)[1]
10,000 randomly selected firms[3]	17.55 (4.12–44.2)[2]
Entire data set	12.37
First half of data set	13.21
Second half of data set	11.40
Lower 10% of rivals	12.32
Upper 10% of rivals	12.45
Lower 25% of rivals	14.32
Upper 25% of rivals	14.20

Notes: [1]Against a general neural network formulation whose number of nodes is determined by the BF. [2]Median value and 99% confidence interval. For the Horowitz [2006] we use the same moment conditions as in our implementation of ELVIS. [3]We implement these tests by randomly sampling 10,000 firms at a time and we do this 1,000 times. We use 60,000 MCMC iterations and we omit the first 10,000 to mitigate possible start up effects.

Appendix A

Table A.1: Variable definitions

Variable	Description and construction of the independent variables
R & D spending (in millions of RMB)	Firm's research and development expenditure. We take the log value of current sales when estimating empirical models
Number of employees	This is the total number of employees in a firm. We take the log value of the number of employees when estimating empirical models. Note that we have deleted observations when the number of employees is reported as zero.
Employee size 1 (<= 50)	This dummy variable identifies the number of employees for a given firm is at least 50. This also represents the first quartile (0 - 25) of the employee distribution and control for firm size.
Employee size 2 (> 50 & <= 95)	This dummy variable identifies the number of employees for a given firm is more than 50 and at least 95. This also represents the second quartile (25-50) of the employee distribution and control for firm size.
Employee size 3 (> 95 & <= 198)	This dummy variable identifies the number of employees for a given firm is more than 95 and at least 198. This also represents the third quartile (50-75) of the employee distribution and control for firm size.
Employee size 4 (> 198)	This dummy variable identifies the number of employees for a given firm is more than 198 and it represents the forth quartile (75-100) of the employee distribution and control for firm size.
Current Sales (RMB million)	This is the firm's total sales in current values (in millions of RMB.) We take the log value of current sales when estimating empirical models. Note that China's National Bureau of Statistics (NBS) survey includes private firms that have sales of over five million RMB. There are no such restrictions on state-owned firms included in the survey. Therefore, we drop all firms with less than five million RMB in sales. We lose 21 observations in this process.
Sales size 1 (<= 17487)	This dummy variable identifies current sales for a given firm is at least 8120 RMB. This also represents the first quartile (0 - 25) of the sales distribution and control for firm size.
Sales size 2 (> 8120 & <= 17487)	This dummy variable identifies current sales for a given firm is more than 8120 RMB and at least 17487 RMB. This also represents the second quartile (25 - 50) of the sales distribution and control for firm size.
Sales size 3 (> 17487 & <= 43690)	This dummy variable identifies current sales for a given firm is more than 17487 RMB and at least 43690 RMB. This also represents the third quartile (50 - 75) of the sales distribution and control for firm size.
Sales size 4 (> 43690)	This dummy variable identifies current sales for a given firm is more than 43690 RMB. This also represents the forth quartile (75 - 100) of the sales distribution and control for firm size.
SIC code (4 digits)	This is the four-digit Standard Industry Classification codes for China.
Number of rival firms	This is the number of rival firms operating in a four-digit SIC code. We take the log value of the number of rival firms [log(number of rivals + 1)] when estimating empirical models
Total profit (in millions of RMB)	We compute each firm's profits by subtracting total intermediate inputs, financial charges, gross wages, amount of current depreciation, and value-added tax from the firm's gross output. A similar method to calculate a firm's corporate profit has been used by Cai and Liu (2009.)
Profit per employee	This is constructed by dividing profits by the number of employees.
Product 1	This dummy variable identifies firms that produce one output.
Product 2	This dummy variable identifies firms that produce two outputs.
Product 3	This dummy variable identifies firms that produce three or more outputs.
Revenue subsidies (in millions of RMB)	A portion of revenue is subsidized by the state government. We take the log value of revenues subsidized [log(revenue subsidized + 1)] when estimating empirical models

Table A.2: Correlations of regression variables

Variable name	Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Log of R & D_{ijlt}	1	1.000														
LI_{ijt}	2	-0.012	1.000													
LI_{ijt}^2	3	-0.024	0.988	1.000												
LI_{ijt}^3	4	-0.029	0.976	0.998	1.000											
Log number of $rivals_{ijlt}$	5	-0.045	-0.094	-0.093	-0.090	1.000										
Log number of $rivals_{ijlt}^2$	6	-0.042	-0.090	-0.089	-0.086	0.926	1.000									
Log number of $rivals_{ijlt}^3$	7	-0.036	-0.081	-0.080	-0.078	0.821	0.971	1.000								
Log of rivals' average R & D_{ijlt}	8	0.173	-0.023	0.028	-0.025	0.374	0.333	0.292	1.000							
Log number of $employees_{it}$	9	0.227	0.034	0.057	0.026	0.040	0.033	0.027	0.049	1.000						
Log of age_{it}	10	0.098	0.056	0.011	0.055	-0.048	-0.038	-0.028	0.016	0.198	1.000					
$Exporter_{it}$	11	0.116	0.006	0.026	0.012	0.106	0.103	0.095	0.072	0.273	0.077	1.000				
Producing $two\ outputs_{it}$	12	0.103	0.023	0.021	0.025	-0.073	-0.066	-0.057	0.032	0.093	0.094	0.051	1.000			
Producing $three\ or\ more\ outputs_{it}$	13	0.108	0.021	0.021	0.019	-0.053	-0.046	-0.039	0.031	0.086	0.079	0.049	0.556	1.000		
Log number of national $universities_{lt}$	14	0.087	0.025	0.022	0.019	-0.022	-0.023	-0.021	0.061	-0.011	0.028	-0.003	0.038	0.045	1.000	
Log number of provincial $universities_{lt}$	15	0.054	0.056	0.058	0.056	-0.007	-0.005	0.000	0.044	0.029	0.036	0.000	0.064	0.048	0.090	1.000

Pearson correlation coefficients

Appendix B

B.1. ELVIS

To explain this method, we define

$$\tilde{g}(y,\theta,\gamma) \equiv \frac{\int g\,(u,y,\theta)\,exp\,\{\gamma' g(u,y,\theta)\}\,d\rho(u|y,\theta)}{\int exp\,\{\gamma' g(u,y,\theta)\}\,d\rho(u|y,\theta)} \tag{B1}$$

where $\rho(u|y,\theta)$ is a user-specified conditional probability measure whose support is $\mathcal{U}$ for all y and θ, and a certain regularity condition is satisfied (Definition 2.2 of Schennach [2014]). The econometrician must specify the measure $\rho(\cdot)$ as input, but any such choice is suitable so long as its support matches the potential support of unobservables and a moment generating function exists and is twice differentiable.[22] Such measures can be independent of θ if stochastic dominance conditions are satisfied for the function $g(u,y,\theta)$. Then, the infinite-dimensional problem of establishing the existence of some measure that solves the original moment conditions is equivalent to the simpler problem of establishing that a finite-dimensional parameter γ solves the following modified moment condition:

$$\inf_{\gamma\in\Re^{d_g}} : ||\mathbb{E}_\pi \tilde{g}(y,\theta,\gamma)|| = 0, \tag{B2}$$

where π denotes that the expectation is with respect to the probability measure of the observable variables. The function $\tilde{g}(y,\theta,\gamma)$ is an average of the original moment condition $g(U,Y,\theta)$ with respect to some distribution of the unobservables that belong to a specific exponential family. It is worth noting that these results require no assumptions, other than measurability, regarding $g(u,y,\theta)$.

To evaluate (B1), given a sample $\{u_{(s)}\}_{s=1}^{S}$ from a distribution proportional to

$$\exp\{\gamma' g(u,y,\theta)\}\,d\rho(u|y,\theta)$$

[22]The choice of $\rho(\cdot)$ has no effect on her results so long as certain properties are satisfied concerning its support and differentiability. Moreover, Schennach [2014] provided a way of constructing $\rho(u|y,\theta)$ automatically to ensure the regularity condition is satisfied; see Proposition 2.1 of her paper.

we use

$$\hat{g}(y,\theta,\gamma) = S^{-1}\sum_{s=1}^{S} g\,(u_{(s)},y,\theta). \tag{B3}$$

Moreover, if the Metropolis–Hastings algorithm is used to obtain the sample, the integrating constant in (B1) need not be known. To construct an average that is a smooth function of θ or γ, Schennach [2014] proposed constructing

$$\tilde{g}(y,\theta,\gamma) = \frac{\int g\,(u,y,\theta)\,exp\,\{\gamma' g(u,y,\theta) - \gamma_o' g(u,y,\theta_o)\}\,r(u|y,\theta_o,\gamma_o)d\rho(u|y,\theta_o)}{\int exp\,\{\gamma' g(u,y,\theta) - \gamma_o' g(u,y,\theta_o)\}\,r(u|y,\theta_o,\gamma_o)d\rho(u|y,\theta_o)}, \tag{B4}$$

where

$$r(u|y,\theta_o,\gamma_o) = \frac{exp\,\{\gamma_o' g(u,y,\theta_o)\}}{\int exp\,\{\gamma_o' g(u,y,\theta_o)\}\,d\rho(u|y,\theta_o)}, \tag{B5}$$

for certain fixed θ_o and γ_o. Then, computing the ratio of averages

$$\hat{g}(y,\theta,\gamma) = \frac{S^{-1}\sum_{s=1}^{S} g\,(u_{(s)},y,\theta)\,exp\,\{\gamma' g(u,y,\theta) - \gamma_o' g(u,y,\theta_o)\}}{S^{-1}\sum_{s=1}^{S} exp\,\{\gamma' g(u,y,\theta) - \gamma_o' g(u,y,\theta_o)\}}. \tag{B6}$$

For given γ_o and θ_o one can evaluate $\tilde{g}(y,\theta,\gamma)$ for all θ and γ by drawing $u_{(s)}$ from a distribution with density proportional to $\exp\{\gamma_o' g(u,y,\theta_o)\}\,d\rho(u|y,\theta_o)$. In turn, averaging over the unobservables then provides us with conventional moment conditions of the form

$$\mathbb{E}\tilde{g}(y,\theta,\gamma) = 0. \tag{B7}$$

B.2. BETEL

Averaging over the unobservables then provides us with conventional moment conditions that involve only observable variables. A viable alternative to fully parametric Bayesian methods, Schennach [2005] suggested BETEL.

To consider this method, suppose we have moment conditions $\{\mathbb{E}[G(y_i,\theta)] =$

$0\}_{i=1}^{n}$. Then, the Bayesian posterior corresponding to BETEL is

$$p(\theta|Y) \propto p(\theta) \prod_{i=1}^{n} w_i^*(\theta), \tag{B8}$$

where $p(\theta)$ is a prior and $\{w_i^*(\theta)\}_{i=1}^{n}$ are solutions to the problem

$$\max_{\{w_i\}_{i=1}^{n}} : -\sum_{i=1}^{n} w_i \log w_i, \tag{B9}$$

subject to the conditions

$$\sum_{i=1}^{n} w_i = 1 \tag{B10}$$

and

$$\sum_{i=1}^{n} w_i G(y_i, \theta) = 0, \tag{B11}$$

provided the interior of the convex hull of $\cup_{i=1}^{n}\{g(y_i, \theta)\}$ contains the origin. The substantial numerical problem is that specifying γ_o and θ_o is not easy, although trial and error can be used. Evaluation of (B1) involves computing

$$\begin{aligned} &\hat{g}(y, \theta, \gamma) \\ &\simeq \frac{S^{-1}\sum_{s=1}^{S} g(u^{(s)}, y, \theta)\, exp\left\{\gamma' g(u^{(s)}, y, \theta) - q(u^{(s)}|y, \alpha)\right\} d\rho(u^{(s)}|y, \theta)}{S^{-1}\sum_{s=1}^{S} exp\left\{\gamma' g(u^{(s)}, y, \theta) - q(u^{(s)}|y, \alpha)\right\} d\rho(u^{(s)}|y, \theta)}, \end{aligned} \tag{B12}$$

provided we can determine an importance density

$$\varphi(u|y, \alpha) = \exp\{Q(u|y, \alpha)\}, \tag{B13}$$

which is "close" to

$$\varpi(u|\theta,\gamma) = \frac{exp\{\gamma' g(u,y,\theta)\}\, d\rho(u|y,\theta)}{\int_u exp\{\gamma' g(u,y,\theta)\}\, d\rho(u|y,\theta)}, \tag{B14}$$

from which a sample $\{u^{(s)}\}_{s=1}^{S}$ can be obtained. If

$$W^{(s)} \equiv \exp\left\{\gamma' g(u^{(s)},y,\theta) - Q(u^{(s)}|y,\alpha)\right\} d\rho(u^{(s)}|y,\theta),$$

it is clear that (B6) reduces to:

$$\begin{aligned}\hat{g}(y,\theta,\gamma) &\simeq \frac{S^{-1}\sum_{s=1}^{S} g\,(u^{(s)},y,\theta)W^{(s)}}{S^{-1}\sum_{s=1}^{S} W^{(s)}}: \\ &= S^{-1}\sum_{s=1}^{S} g\,(u^{(s)},y,\theta)w^{(s)},\end{aligned} \tag{B15}$$

where $w^{(s)} = \frac{W^{(s)}}{S^{-1}\sum_{s'=1}^{S} W^{(s')}}$, $s = 1,\ldots,S$, which is trivial to compute. What is not trivial is to determine an importance density $\varphi(u|y,\alpha)$, which by choice of parameters $\alpha \in \mathcal{A} \subseteq \Re^{d_\alpha}$ can approximate (B14) across all values of u, θ and γ; see Danielsson and Richard [1993].

B.3. Markov Chain Monte Carlo

We used the algorithm proposed by Girolami and Calderhead [2011] (GC) to update draws for $\boldsymbol{\theta}$. The algorithm uses local information about both the gradient and the Hessian of the log-posterior distribution conditional on $\boldsymbol{\theta}$ at the existing draw. A Metropolis test is used for accepting the candidate generated, but the GC algorithm moves considerably faster relative to alternatives. The GC algorithm is started at the first-stage GMM estimator and MCMC is run until convergence. Depending on the model and the subsample this takes 5,000 to 10,000 iterations. For safety we run 10,000 iterations. Then we run another 150,000 MCMC iterations omitting the first 50,000 iterations to obtain final results for posterior moments, densities of parameters, and functions of interest. It has been found that the GC algorithm performs vastly superior relative to the standard Metropolis–Hastings algorithm and autocorrelations are much smaller.

For convenience, let $L(\boldsymbol{\theta}) = \log\, p\,(\boldsymbol{\theta}|\boldsymbol{X})$ denote the log posterior of $\boldsymbol{\theta}$. Moreover, define

$$\boldsymbol{G}(\boldsymbol{\theta}) \equiv \text{est. cov}\frac{\partial}{\partial \boldsymbol{\theta}}\log\, p\,(\boldsymbol{X}|\boldsymbol{\theta}) \tag{B16}$$

to be the empirical counterpart of

$$\boldsymbol{G}_o(\boldsymbol{\theta}) = -E_{Y|\boldsymbol{\theta}}\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\log\, p\,(\boldsymbol{X}|\boldsymbol{\theta}). \tag{B17}$$

The Langevin diffusion is given by the following stochastic differential equation

$$d\boldsymbol{\theta}(t) = \frac{1}{2}\tilde{\nabla}_{\boldsymbol{\theta}} L\{\boldsymbol{\theta}(t)\}dt + d\boldsymbol{B}(t) \tag{B18}$$

where

$$\tilde{\nabla}_{\boldsymbol{\theta}} L\{\boldsymbol{\theta}(t)\} = -\boldsymbol{G}^{-1}\{\boldsymbol{\theta}(t)\}\cdot \tilde{\nabla}_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\} \tag{B19}$$

is the so called "natural gradient" of the Riemann manifold generated by the log posterior.
The elements of the Brownian motion are

$$\begin{aligned}
&\boldsymbol{G}^{-1}\{\boldsymbol{\theta}(t)\}d\boldsymbol{B}_i(t)\\
&= |\boldsymbol{G}\{\boldsymbol{\theta}(t)\}|^{-\frac{1}{2}}\sum_{j=1}^{K_\beta}\frac{\partial}{\partial \boldsymbol{\theta}}\left[\boldsymbol{G}^{-1}\{\boldsymbol{\theta}(t)\}_{ij}|\boldsymbol{G}\{\boldsymbol{\theta}(t)\}|^{\frac{1}{2}}\right]dt\\
&+ \left[\sqrt{\boldsymbol{G}\{\boldsymbol{\theta}(t)\}}d\boldsymbol{B}(t)\right]_i.
\end{aligned} \tag{B20}$$

The discrete form of the stochastic differential equation provides a proposal as follows:

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i^o &+ \frac{\varepsilon^2}{2}\{\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^o)\}_i\\
&+ \frac{\varepsilon^2}{2}\sum_{j=1}^{K_\theta}\{\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)\}_{ij}\,\text{tr}\left\{\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)\frac{\partial \boldsymbol{G}(\boldsymbol{a}^o)}{\partial \boldsymbol{\theta}_j}\right\}\\
&- \varepsilon^2\sum_{j=1}^{K_\theta}\left\{\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)\frac{\partial \boldsymbol{G}(\boldsymbol{\theta}^o)}{\partial \boldsymbol{\theta}_j}\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)\right\}_{ij}\\
&+ \left\{\varepsilon\sqrt{\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)}\boldsymbol{\xi}^o\right\}_i\\
&= \boldsymbol{\mu}(\boldsymbol{\theta}^o,\varepsilon)_i + \left\{\varepsilon\sqrt{\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)}\boldsymbol{\xi}^o\right\}_i
\end{aligned}$$

where $\boldsymbol{\beta}^o$ is the current draw. The proposal density is

$$q\left(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^o\right) = N_{K_\theta}\left(\tilde{\boldsymbol{\theta}}, \varepsilon^2\boldsymbol{G}^{-1}(\boldsymbol{\theta}^o)\right) \tag{B21}$$

and convergence to the invariant distribution is ensured by using the standard form Metropolis–Hastings probability

$$\min\left\{1, \frac{p\left(\widetilde{\boldsymbol{\theta}}|\cdot, Y\right)q\left(\boldsymbol{\theta}^{o}|\widetilde{\boldsymbol{\theta}}\right)}{p\left(\boldsymbol{\theta}^{o}|\cdot, Y\right)q\left(\widetilde{\boldsymbol{\theta}}|\boldsymbol{a}^{o}\right)}\right\}. \tag{B22}$$

Finally, $\boldsymbol{\varepsilon}$ is selected so that the acceptance rate of the algorithm is not too small or large. We determined $\boldsymbol{\varepsilon}$ using a target acceptance rate of, approximately, 25%.