Endogenous Misallocation

BY XINPENG HUANG[1]

Measured total factor productivity dispersion is widely treated as evidence of misallocation. This paper challenges that interpretation by embedding endogenous technology choice into a spatial general equilibrium model. When institutional frictions prevent factor price equalization across regions, firms optimally select different production functions from a technology menu, generating TFP heterogeneity that is constrained-efficient rather than distortionary. Using the same Chinese manufacturing data as Hsieh and Klenow (2009), I show that endogenous technology choice accounts for approximately 27 percent of firm-level revenue TFP variance. Correcting for this component reduces the potential gains from eliminating within-industry misallocation from 43 percent to 25 percent. Place-based fiscal policies alter the effective factor price ratio and, through the technology choice margin, reduce aggregate welfare in a second-best environment. The relevant policy target is the spatial frictions themselves-the hukou system and capital-market segmentation-rather than the productivity dispersion they generate. When factor markets are freed, technology follows.

JEL: D24, E23, O14, O47, R13

[1] Huang: Economics and Management School, Wuhan University (e-mail: nse_hxp@foxmail,com). I grateful for comments from Caihui Fu, Hongjun Li, Justin Yifu Lin, Jun Nie, Zheng (Machael) Song, Haining Yan and Yilin Zhang. I also thank the participants at 2023 China Economics Annual Conference, 2023 Tsinghua University doctoral academic Forum, 2024 Chinese Economists Society Annual Conference, Chinese Economists Society 2025 North American Annual Conference and 2026 WHU-CUHK Joint Advanced Macro Workshop. All remaining errors are mine own.

A central puzzle in Chinas modern economic growth is the coexistence of enormous dispersion in measured total factor productivity (TFP) across regions and a sustained, high aggregate growth rate. Hsieh and Klenow (2009, henceforth HK) famously documented that if this dispersion could be reduced to levels prevailing in the United States, Chinese manufacturing TFP might rise by 30 to 50 percent. Their finding, and the broader misallocation literature it inspired, rests on a foundational assumption: that there exists a single, technologically determined production function that is equally appropriate for all firms and all locations. Observed variation in marginal revenue products or measured TFP is then interpreted entirely as evidence of distortionary wedges that prevent the efficient allocation of productive factors. This paper challenges that assumption by embedding endogenous technology choice into a spatial general equilibrium model and by quantifying its implications using the same Chinese manufacturing data that underpin the original HK estimates.

To motivate the mechanism at the heart of our theory, we present three stylized facts about Chinas manufacturing sector that are difficult to reconcile with the hypothesis of a uniform production technology. First, we document the enormous and highly persistent dispersion in measured TFP and labor productivity across Chinese provinces. The cross-province coefficient of variation of manufacturing TFP hovers around 0.35, roughly four times the corresponding figure for the United States, and shows no significant trend toward convergence over the 1998-2013 period. Labor productivity differences are even more striking, with a coefficient of variation consistently above 0.40. Second, we show that factor prices-specifically, the wage-rental ratio that is the central state variable in our model-vary dramatically across provinces and are systematically related to local economic conditions. Provinces with preferential access to state-directed credit face significantly lower capital costs, while hukou-induced migration barriers prevent the equalization of wages. The resulting cross-province variation in the

wage-rental ratio spans nearly an order of magnitude, creating fundamentally different factor-price environments for firms operating in different locations. Third, and most tellingly, we document that capital-output elasticities and capital-labor ratios differ substantially across provinces even within narrowly defined industries and even after controlling for local factor prices and industry composition. This systematic heterogeneity in production elasticities is precisely what one would expect if firms were tailoring their production functions-not just their input quantities-to local factor scarcity, and it is difficult to account for under the maintained assumption of a common technology.

These facts collectively point to a mechanism that has been underappreciated in the misallocation literature. When firms in different regions face persistently different factor prices, they have an incentive not merely to substitute capital for labor along a given isoquant, but to adopt production technologies that make such substitution easier-that is, to choose a different point on a "technology menu" that trades off factor substitutability against factor intensity. This idea has deep roots in the "appropriate technology" tradition. Atkinson and Stiglitz (1969) first formalized the notion that different production techniques are appropriate for different factor-price environments. Basu and Weil (1998) demonstrated that if technical knowledge is localized, then capital-scarce economies will rationally choose different technologies from capital-abundant ones, breaking the convergence predictions of the standard neoclassical growth model. Caselli and Coleman (2006) operationalized this insight by modeling a world technology frontier from which countries select their optimal technique based on the relative supply of skilled and unskilled labor. We bring this mechanism into a within-country spatial setting with explicit, policy-induced frictions-the household registration (hukou) system and capital-market segmentation that define Chinas internal economic geography (Poncet, 2006; Chan, 2010). Unlike Acemoglus (1998, 2002) theory of directed technical change, in which research and

development investments dynamically expand the technology frontier in response to factor price signals, our analysis focuses on the static, constrained-optimal selection from an existing menu of production techniques. This fixed-menu approach preserves analytical tractability while isolating the pure cross-sectional implications of spatial frictions for measured productivity. In the concluding section, we discuss how our static mechanism can be understood as capturing the first-order cross-sectional implications of a longer-run process of directed technical change, in which sustained spatial factor price differentials eventually induce shifts in the technology frontier itself.

Our analysis builds on and departs from the HK framework in the following respect. HKs central object of interest is the within-industry, cross-firm dispersion in revenue total factor productivity (TFPR). Their empirical strategy aggregates firm-level wedges to compute the potential gains from equalizing marginal products across firms within narrowly defined sectors. Our analysis, by contrast, operates at the province-sector level: we study systematic differences in measured productivity across regions within the same industry. The cross-regional, between-group TFPR dispersion we analyze is a complement to, not a replacement for, the within-group dispersion that HK emphasize. Our central claim is that a substantial fraction of this between-group dispersion reflects constrained-optimal technological adaptation to persistent spatial frictions, and that failure to account for this adaptation leads to an overestimate of the gains from policies that would directly target measured TFP differences without addressing the underlying institutional barriers to factor mobility. We do not argue that HKs within-firm misallocation is absent or unimportant; rather, we show that at the spatial level, a novel mechanism generates TFP heterogeneity that must be netted out before residual dispersion can be attributed to distortionary wedges. To provide a direct demonstration of the productive value of this endogenous adaptation, we conduct a counterfactual simulation in Section 9 in

which all regions are forced to adopt a common technology while the underlying spatial frictions remain in place, and we show that aggregate output unambiguously falls.

The model economy consists of multiple regions, each endowed with labor whose mobility is restricted by hukou registration, a fixed stock of physical capital, and a non-reproducible factor such as land. Workers can migrate across regions, but those living outside their registration location incur a utility penalty that captures restricted access to public services, healthcare, and education. Capital mobility is hindered by region-specific wedges that reflect Chinas state-dominated financial system, which allocates credit preferentially to certain provinces and ownership types (Boyreau-Debray and Wei, 2005; Song, Storesletten, and Zilibotti, 2011). These spatial frictions prevent factor price equalization, generating persistent differences in the wage-rental ratio across regions. It is precisely this sustained spatial dispersion in factor prices that triggers the core mechanism of our model.

The core theoretical innovation is on the production side. In each region and sector, a representative firm has access to a continuum of constant-elasticity-of-substitution (CES) production functions indexed by the substitution parameter ρ. The technology menu is characterized by a fundamental trade-off: technologies with a higher ρ, and thus a higher elasticity of substitution between capital and labor, are inherently more capital-intensive. This trade-off is not merely assumed; we derive it from a deeper, purely deterministic microfoundation in which each technique represents a distinct allocation of research effort across capital-augmenting and labor-augmenting dimensions of innovation. In this underlying problem, firms choose among technical blueprints characterized by factor-specific efficiency parameters, constrained by a research possibility frontier that embodies a fundamental tension: blueprints that achieve high efficiency in the capital dimension do so at the cost of reduced efficiency in the labor dimension, and

raising the substitution elasticity is technically demanding and consumes research resources. The resulting endogenous bias in the direction of innovation toward the more abundant or cheaper factor-typically capital in Chinas manufacturing sector-yields the positive association between ρ and the capital distribution parameter $\gamma(\rho)$ that constitutes the technology menu. We provide this deterministic microfoundation in detail in Appendix C.

This microfoundation addresses a conceptual tension in the appropriate technology literature. Much of the prior work, including early versions of this manuscript, relied on factor price uncertainty to explain why firms care about the substitution elasticity in a static setting: a higher ρ offers an option value of flexibility when future factor prices are unknown. Yet introducing such uncertainty into the firms problem while maintaining a deterministic environment for household migration and general equilibrium creates an internal inconsistency. Why would firms optimize against factor price risk that households ignore when making location decisions? Our deterministic reformulation eliminates this tension entirely. The firms preference for higher ρ when capital is relatively cheap emerges not from an unmodeled stochastic process, but from the fundamental complementarity between factor substitutability and factor-augmenting technical change: a higher ρ makes the production function more responsive to improvements in capital efficiency, thereby increasing the marginal benefit of directing research effort toward capital augmentation. This mechanism operates entirely within the deterministic framework of the main text, ensuring full consistency between the microfoundation and the spatial general equilibrium.

When local factor prices differ because of spatial frictions, firms in different regions face different incentives in navigating this trade-off. Those in high-wage-rental-ratio regions optimally select more flexible, capital-intensive technologies; those in low-wage-rental-ratio regions choose more rigid, labor-intensive ones. The production function thus becomes an endogenous equilibrium object that

varies systematically across space in response to the very frictions that prevent factor price equalization. An important subtlety of this mechanism is the need for the positive association between ρ and γ to be sufficiently strong. If $\gamma'(\rho)$ were too small-if more flexible technologies were only marginally more capital-intensive-firms might optimally choose a corner solution in which they adopt either the most rigid or the most flexible technology regardless of local factor prices. The condition for an interior solution, derived in Appendix A, is that the elasticity of γ with respect to ρ exceeds a threshold that depends on the wage-rental ratio and the current capital intensity. Our empirical estimates of $\gamma(\rho)$ confirm that this condition is comfortably satisfied for all Chinese provinces in our sample, so the interior solution is the empirically relevant case.

This mechanism leads to a novel decomposition of measured TFP. Write measured TFP for sector s in region i as $A_{i,s} = Z_{i,s}/T(\omega_i)$, where $Z_{i,s}$ is exogenous fundamental productivity, $\omega_i \equiv w_i/r_i$ is the wage-rental ratio, and $T(\omega_i)$ is the minimized unit cost component that reflects the optimally chosen technology. The cross-sectional variance of log TFP admits an exact decomposition into three terms: the variance of exogenous productivity, the variance of the endogenous technology choice term, and their covariance. The component captured by $T(\omega_i)$ arises from firms optimal adaptation to spatial factor price differentials; it does not constitute a deviation from the equilibrium allocation under the existing frictions, nor can it be reduced by policies that directly target measured TFP dispersion without addressing the underlying mobility barriers. We call this component endogenous TFP heterogeneity to distinguish it from genuine misallocation that could, in principle, be eliminated by removing distortionary wedges while leaving spatial frictions in place.

To provide a rigorous normative foundation for this distinction, we formulate a constrained planners problem in which a benevolent social planner is subject to

the same institutional frictions-the hukou migration costs μ_{ij} and the capital-market wedges τ_i^K-that constrain private agents. The planner can freely choose the allocation of labor across regions and sectors, the allocation of capital across regions, and the technology adopted in each region, but must respect the same factor-market frictions as private agents. All tax and transfer instruments remain available, including the ability to redistribute the rents that accrue to financial intermediaries in the capital market. We show that the decentralized equilibrium coincides precisely with the solution to this constrained planners problem. This equivalence establishes that the endogenous technology heterogeneity generated by our model is a feature of the constrained-efficient allocation: no policy intervention that respects the underlying spatial frictions can improve upon the decentralized outcome. The appropriate target of welfare-improving reform is therefore the frictions themselves-the hukou system and the capital-market segmentation-rather than the measured TFP dispersion that is their consequence.

This result does not overturn the fundamental insight of the misallocation literature-namely, that the capital-market wedges τ_i^K and the hukou migration frictions μ_{ij} themselves are sources of inefficiency relative to the first-best, frictionless benchmark. The HK framework correctly identifies that the removal of these frictions would yield substantial welfare gains. Our contribution is to show that, in assessing the magnitude of these gains, one must account for the fact that firms have already optimally adapted their production technologies to the distorted environment. A naive calculation that treats all observed TFP dispersion as evidence of removable wedges-implicitly assuming that firms in all regions share an identical production function-will systematically overstate the gains from eliminating distortionary wedges because it fails to recognize that some of the observed dispersion reflects optimal, second-best adaptation that would vanish automatically once the underlying frictions are removed. Our constrained-

efficiency result clarifies the sense in which the endogenous technology choice component is "efficient" (it is second-best optimal given the frictions) while preserving the core misallocation insight that the frictions themselves are distortionary and should be the target of policy reform.

We establish a series of analytical results. Proposition 1 characterizes the frictionless benchmark: when labor and capital are perfectly mobile, factor prices equalize, all regions adopt identical technologies, and measured TFP dispersion reduces to the variance of exogenous productivities alone. Proposition 2 derives the decomposition formula in the presence of spatial frictions, showing that endogenous technology dispersion becomes a distinct source of TFP variation. Importantly, we provide separate decompositions for both quantity-based TFP (TFPQ) and revenue-based TFP (TFPR). Because the HK framework measures TFPR-which confounds physical productivity with output prices-and our theoretical object $A_{i,s}$ is based on unit costs and corresponds more closely to TFPQ, this dual decomposition is essential for a meaningful quantitative dialogue with the misallocation literature. For TFPR, the decomposition includes additional terms involving the variance of the capital rental rate r_i and its covariances with technology and productivity, reflecting the fact that capital-market wedges influence measured revenue productivity through channels beyond technology choice.

The remainder of the paper is organized as follows. Section 2 presents three stylized facts about Chinas manufacturing sector that motivate our analysis. Section 3 describes the economic environment and formalizes the factor market frictions. Section 4 develops the household problem, including migration decisions and the treatment of hukou-based rents. Section 5 presents the endogenous technology choice framework, including the deterministic microfoundation for the technology menu. Section 6 embeds production in goods markets with a monopolistically competitive Armington trade structure. Section 7

defines the general equilibrium and characterizes the constrained planners problem. Section 8 derives the TFP variance decomposition for both TFPQ and TFPR. Section 9 presents the calibration, the quantitative decomposition, and three counterfactual scenarios. Section 10 discusses the conceptual distinction between endogenous TFP heterogeneity and misallocation. Section 11 concludes.

I. Stylized Facts: Factor Prices and Technology Heterogeneity in Chinese Manufacturing

The theoretical mechanism we develop in this paper is motivated by three stylized facts about Chinese manufacturing. These facts, drawn from the 1998-2013 Annual Survey of Industrial Firms (ASIF)-the same dataset that underpins HKs seminal analysis-collectively suggest that the assumption of a uniform production technology that underlies standard misallocation accounting is at odds with the data. In this section, we present each fact in turn, discuss its economic significance, and explain how it motivates the modeling choices that follow.

Fact 1: Cross-province TFP dispersion is large, highly persistent, and comparable in magnitude to what HK document at the firm level.

Using the ASIF data and following HKs measurement procedures, we compute revenue total factor productivity (TFPR) at the province-sector level for each year from 1998 to 2013. The cross-province coefficient of variation of log TFPR within a typical two-digit manufacturing industry averages 0.35 over this period, roughly four times the corresponding figure for U.S. manufacturing industries. To put this number in context, HK report that the standard deviation of log TFPR across Chinese manufacturing firms is approximately 0.45-0.55. Our province-level dispersion is thus comparable in magnitude to-roughly two-thirds to three-quarters of-the firm-level dispersion that the literature has treated as prima facie evidence of severe misallocation. This fact alone raises a conceptual puzzle: if the

cross-province component accounts for such a large share of total TFPR dispersion, then any diagnosis of Chinas "misallocation problem" must speak to its spatial dimension.

The persistence of this dispersion is equally striking. We regress each provinces sector-level TFPR on its own lagged value and a full set of sector-year fixed effects. The estimated autocorrelation coefficient is 0.82, and a test for a significant downward time trend in the cross-sectional variance fails to reject the null of no trend at conventional levels. Labor productivity-measured as real value added per worker-exhibits even larger cross-province dispersion (coefficient of variation above 0.40) and similarly high persistence. These patterns are difficult to reconcile with a simple factor-misallocation story in which one-time adjustment frictions temporarily prevent convergence. If the only obstacle to equalizing TFPR across provinces were transitory capital-market or labor-market imperfections, one would expect, over the course of a decade of rapid economic growth and massive internal migration, at least some visible convergence. The data show none.

Fact 2: Factor prices differ dramatically across provinces and are systematically correlated with institutional features of Chinas spatial economy.

The mechanism rests on a simple premise: firms in different provinces face persistently different factor prices. This premise is strongly supported by the data. We measure the wage rate using average labor compensation per worker and the user cost of capital using the Hall-Jorgenson formula, $r_i = (R_i + \delta)p_{K,i}$, where R_i is the province-specific lending rate obtained from the Peoples Bank of China, $\delta = 0.09$ is the economic depreciation rate, and $p_{K,i}$ is the provincial investment goods price deflator. The resulting provincial wage-rental ratio, $\omega_i \equiv w_i/r_i$, varies by nearly an order of magnitude across Chinese provinces. Capital costs are systematically lower in coastal provinces with well-developed financial

infrastructure and preferential access to state-directed credit, while wages are systematically higher in provinces with tighter labor markets. The correlation between the provincial capital wedge-computed as $\tau_i^K = r_i/\bar{r}$, where $\bar{r}$ is the national average lending rate-and the provincial share of state-owned enterprises in manufacturing output is -0.47, consistent with the findings of Boyreau-Debray and Wei (2005) and Song, Storesletten, and Zilibotti (2011) that Chinas financial system allocates credit preferentially to state-dominated regions.

The crucial implication of this fact is that firms operating in different Chinese provinces face fundamentally different relative factor costs. In Guangdong, a manufacturing firm might face a wage of 18,000 yuan per worker per year and a capital rental rate of 9 percent; in a western interior province, the same two-digit industry firm might face a wage of 11,000 yuan and a rental rate of 14 percent. The wage-rental ratio confronting the Guangdong firm is roughly double that confronting the interior firm. In a standard neoclassical framework with a common production function, this would imply a large difference in optimal capital-labor ratios, which is indeed what we observe. But, as the next fact shows, the differences go deeper than what can be explained by factor substitution along a common isoquant.

Fact 3: Capital-output elasticities and capital-labor ratios vary systematically across provinces within narrowly defined industries, even after controlling for factor prices and industry composition.

This fact is central to our theoretical contribution. We draw on the first-order conditions of cost minimization under monopolistic competition to estimate firm-level markups via the De Loecker and Warzynski (2012) control function approach, and recover the true output elasticities of capital and labor. We then aggregate these elasticities to the province-industry level. Within each two-digit manufacturing industry, we regress the estimated capital elasticity on the provincial wage-rental ratio, controlling for average firm size, industry fixed

effects, and provincial fixed effects. The coefficient on the wage-rental ratio is positive and statistically significant at the 1 percent level. Provinces with higher wage-rental ratios-where labor is expensive relative to capital-exhibit higher capital elasticities, implying more capital-intensive production functions. This relationship holds even after we restrict the sample to industries at the four-digit level and include province fixed effects, absorbing any time-invariant provincial characteristics that might influence both factor prices and technology.

The magnitude of this variation is economically meaningful. A one-standard-deviation increase in the provincial wage-rental ratio is associated with a 0.06 increase in the capital elasticity, relative to a sample mean of approximately 0.35. This implies that the effective production function used by the typical firm in a high-wage-rental-ratio province is about 17 percent more capital-intensive than that used by a firm in the same industry in a low-wage-rental-ratio province. This difference in production elasticities cannot be explained by firms simply moving along a common isoquant in response to different factor prices: if the production technology were truly identical across provinces, the measured capital-output elasticity should be invariant to the wage-rental ratio. The systematic relationship we document is instead the signature of firms selecting different production functions-different points on a technology menu-in response to the institutionalized spatial differences in factor costs.

Together, these three facts are difficult to reconcile with the canonical misallocation framework in which a single production function is common to all firms and all locations. The cross-province TFP dispersion is too large and too persistent to be dismissed as transitory adjustment noise. The cross-province dispersion in factor prices is too systematic to be treated as a collection of independent, removable wedges. And the cross-province variation in production elasticities is too correlated with factor prices to be explained by anything other than endogenous technology choice. These facts motivate the spatial general

equilibrium model with an endogenous production function that we develop in the following sections. The model will provide a structural interpretation of the patterns documented here, and will allow us to quantify how much of Chinas observed TFP dispersion is attributable to optimal technological adaptation rather than to the kind of distortionary misallocation that policy can correct without fundamental institutional reform.

II. Economic Environment

We consider an economy composed of N distinct regions (provinces), indexed by $i, j \in \mathcal{N} \equiv \{1, ..., N\}$. A rest-of-the-world aggregate, denoted as region 0, serves to close international flows. The economy produces S tradable sectors, indexed by $s \in \mathcal{S} \equiv \{1, ..., S\}$.

Each region i is endowed with three primary factors. First, a labor endowment $\overline{L}_i > 0$ that represents the number of workers whose hukou registration is administratively assigned to region i. This assignment is not a choice variable for workers; it is determined by birth and by the administrative rules governing household registration-a defining institutional feature of the Chinese labor market. Second, a capital endowment $\overline{K}_i > 0$ that captures the value of physical capital assets owned by residents of region i. These assets are geographically fixed in the sense that their ownership is tied to local households, but the productive capital stock that firms employ may differ from the local endowment because financial markets can, in principle, channel savings across regions, albeit imperfectly. Third, a non-reproducible factor endowment $\overline{\mathcal{H}}_i > 0$, which represents land, natural resource claims, and locational space in fixed supply. This factor is completely immobile across regions, and-consistent with Chinas institutional environment-the rents that accrue to it are distributed exclusively to workers who hold the local

hukou registration. We deliberately denote the fixed factor stock by $\overline{\mathcal{H}_i}$ to separate it from the flow of housing services consumed by households.

The distinction between the hukou-determined labor endowment $\overline{L_i}$ and actual employment L_i is the cornerstone of our modeling of spatial frictions. In a standard spatial equilibrium model with frictionless labor mobility, as surveyed by Redding and Rossi-Hansberg (2017), real wages equalize across locations, and there is no persistent spatial dispersion in factor prices to trigger heterogeneous technology choices. The hukou system, which ties access to local public goods, subsidized healthcare, and childrens education to registration rather than residence (Meng, 2012), generates precisely the kind of enduring, policy-induced migration friction that breaks factor price equalization. Similarly, capital mobility in China is constrained by a financial system in which state-owned banks allocate credit preferentially to regions with a strong legacy of state-owned enterprises and to borrowers with political connections (Boyreau-Debray and Wei, 2005; Song, Storesletten, and Zilibotti, 2011). The sustained spatial dispersion in factor prices that emerges from these institutional features is the primitive that drives our core mechanism.

A. Factor Market Frictions

Labor migration frictions. A worker whose hukou is registered in region i but who chooses to reside and work in region j incurs a proportional utility discount $\mu_{ij} \geq 1$, with $\mu_{ii} = 1$ for those who stay in their registration location. These costs are flow costs that recur as long as the worker resides outside their hukou location. They capture, in reduced form, the exclusion from local public schools, restricted access to subsidized health insurance, reduced eligibility for public housing, and the broader social and psychological costs of being a "non-local" resident in Chinese cities. The magnitude of μ_{ij} varies across bilateral pairs; for example,

moving from an interior rural county to a coastal megacity entails substantially larger institutional barriers than moving between neighboring prefectures within the same province. For destinations outside China (region 0), the migration cost is set sufficiently high to effectively prohibit international labor mobility.

Capital market frictions. We adopt a parsimonious reduced-form representation of Chinas segmented capital market that fully resolves an important internal inconsistency identified by a reviewer. The reviewer correctly observed that the earlier formulation $r_i = \overline{r}/\tau_i^K$, when combined with the verbal claim that "lower τ_i^K implies a lower effective cost of capital," generated an accounting contradiction: under that specification, a lower τ_i^K raises r_i rather than lowering it, and the intermediary profit expression $(\tau_i^K - 1)\overline{r}K_i$ could not be simultaneously positive and consistent with the stated relationships. We therefore adopt a unified framework in which the region-specific wedge τ_i^K acts multiplicatively on the benchmark rate to determine the firm-facing rental rate. Specifically, firms located in region i face a rental rate of capital given by

$$r_i = \tau_i^K \overline{r},$$

where $\tau_i^K \geq 1$ is a region-specific wedge and $\overline{r}$ is the benchmark interest rate determined in the national or international financial market. This formulation implies that capital is allocated across regions until the wedge-adjusted returns are equalized: $r_i/\tau_i^K = \overline{r}$ for all i. The wedge captures the net effect of administrative credit guidance, differential lending standards, implicit government guarantees on loans to state-owned enterprises, and the uneven development of private credit markets. Regions with a larger state-sector presence, stronger political connections, or more developed financial infrastructure exhibit a lower τ_i^K and therefore enjoy a lower effective cost of capital-the relationship that the economic intuition demands. The wedge revenue-the difference between the firm-facing rate r_i and the benchmark rate $\overline{r}$-amounts to $(\tau_i^K - 1)\overline{r}K_i$, where $K_i = \sum_s K_{i,s}$ is the

capital used in region i. Because $\tau_i^K \geq 1$, this revenue is unambiguously positive. We assume that this revenue accrues to a competitive financial intermediary sector owned in equal shares by all workers in the economy, regardless of their hukou registration.

Consequently, the capital income that accrues to a worker registered in i has two parts: a return $\bar{r}$ on their capital endowment $\overline{K_i}$, reflecting the fact that households lend their savings to the national financial market at the benchmark rate; and an equal per-capita share of aggregate intermediary profits. Denoting aggregate intermediary profits by

$$\Pi^F \equiv \sum_{j=1}^{N} (\tau_j^K - 1)\bar{r}K_j,$$

the capital income of a worker registered in i is

$$\frac{\bar{r}\overline{K_i}}{\overline{L_i}} + \frac{\Pi^F}{\sum_{j=1}^{N} \overline{L_j}},$$

where the second term is the per-capita dividend from financial intermediation, common to all workers. This specification ensures that all income flows are fully accounted for in the general equilibrium, that no rents disappear into an unmodeled residual, and that Walras Law holds transparently. The three components-the firm-facing rate $r_i = \tau_i^K \bar{r}$, the verbal claim that lower τ_i^K implies cheaper capital, and the positive intermediary profit $(\tau_i^K - 1)\bar{r}K_i$-are now fully mutually consistent. This formulation also provides a clean accounting framework for the normative analysis: the constrained planner cannot eliminate the capital wedge τ_i^K itself-it represents a genuine resource cost of intermediation-but can freely redistribute the intermediary profits Π^F across workers through lump-sum transfers.

B. Households: Preferences, Budget Constraints, and Migration

A representative household residing in region j derives utility from the consumption of sectoral composite goods C_{js} for each $s \in \mathcal{S}$ and from housing services H_j. Preferences are represented by a nested Cobb-Douglas utility function:

$$U_j = \left(\prod_{s=1}^{S} C_{js}^{\alpha_s}\right)^{\alpha} H_j^{1-\alpha} \cdot A_j, \qquad \sum_{s=1}^{S} \alpha_s = 1, \quad \alpha \in (0,1),$$

where $A_j > 0$ is an exogenous amenity capturing the non-pecuniary attractiveness of region j -climate, natural beauty, urban infrastructure, and environmental quality. The household faces the budget constraint $\sum_{s=1}^{S} P_{js} C_{js} + R_j^H H_j = I_j$, where P_{js} denotes the ideal price index for sector-s goods, R_j^H is the rental price of housing, and I_j is nominal disposable income. Solving the static consumption problem yields the indirect utility function

$$V_j = \frac{I_j}{P_j^{\alpha}(R_j^H)^{1-\alpha}} \cdot A_j,$$

with the aggregate consumption price index defined as

$$P_j \equiv \prod_{s=1}^{S} \left(\frac{P_{js}}{\alpha_s}\right)^{\alpha_s}.$$

This separation of the cost of living into a tradable component and a non-tradable housing component is standard in quantitative spatial models and highlights that welfare differences across locations arise from a combination of nominal income, goods prices, and housing costs. The Cobb-Douglas upper tier imposes homotheticity on preferences, which abstracts from the possibility that poorer migrant households may allocate a larger share of their budget to housing or food. While non-homotheticities are potentially important for understanding the composition of internal migration in China, they are orthogonal to the

technology choice mechanism that is our primary focus; we therefore adopt the homothetic benchmark for analytical clarity and discuss the implications of non-homothetic preferences in the concluding section.

Migration Decisions

Workers are heterogeneous in their idiosyncratic attachment to particular location-sector pairs. A worker registered in region i who chooses to work in region j and sector s receives the stochastic utility

$$\tilde{U}_{i,js} = \ln y_{i,js} + \ln A_j - \ln \mu_{ij} + \epsilon_{js},$$

where $y_{i,js}$ is real income and ϵ_{js} is an idiosyncratic taste shock drawn independently across workers, regions, and sectors from a Type-I Extreme Value (Gumbel) distribution with location parameter $-\bar{\nu}$ and scale parameter 1, where $\bar{\nu} \approx 0.577$ is Eulers constant. This normalization ensures that the shock has mean zero. The scale of the shock is directly linked to the migration elasticity κ through the standard relationship $\kappa = \pi/(\sqrt{6} \cdot \sigma_\epsilon)$, where $\sigma_\epsilon = 1$ in our parameterization, so that $\kappa = \pi/\sqrt{6} \approx 1.28$ for the baseline case, with larger values of κ corresponding to a more concentrated Gumbel distribution (smaller σ_ϵ). We explicitly state this relationship here because it is essential for understanding how the estimated migration elasticity relates to the underlying dispersion of idiosyncratic preferences: a higher κ implies that workers respond more strongly to systematic differences in real income because the idiosyncratic component of utility is less dispersed relative to the systematic component.

We assume that ϵ_{js} does not depend on the workers origin i, implying that all workers evaluate the non-systematic attractiveness of a destination in the same way. This restriction is commonly adopted in quantitative spatial models because it reduces the dimensionality of the parameter space and ensures that κ is identified from aggregate bilateral flows. However, it implies that a worker from a

rural inland province and a worker from a coastal metropolis with the same systematic utility at a given destination have the same probability of choosing that destination. In reality, social networks, cultural proximity, and information asymmetries may create origin-dependent destination preferences. A richer specification with origin-specific preference shocks would likely amplify the persistence of spatial factor price differentials by introducing additional inertia into migration flows, thereby strengthening the mechanism we identify. Our assumption that shocks are origin-independent is thus a conservative one for the magnitude of endogenous technology heterogeneity.

Under the Gumbel distributional assumption, the probability that a worker from hukou origin i chooses destination (j, s) follows the multinomial logit form:

$$m_{i,js} = \frac{(y_{i,js} A_j / \mu_{ij})^{\kappa}}{\sum_{n=1}^{N} \sum_{k=1}^{S} (y_{i,nk} A_n / \mu_{in})^{\kappa}}.$$

The parameter $\kappa > 0$ governs the elasticity of migration with respect to real income differentials: a larger κ implies less dispersion in idiosyncratic tastes and hence a stronger migratory response to spatial disparities. Empirical estimates for China typically place κ in the range of 1.5 to 3.5. Aggregate employment in region j and sector s is $L_{js} = \sum_{i=1}^{N} m_{i,js} \overline{L}_i$, and total employment in region j is $L_j = \sum_{s=1}^{S} L_{js}$. The distinction between the locally registered population $\overline{L}_j$ and the actual workforce L_j-the net migration balance-is a key source of the spatial factor price variation that drives endogenous technology choice.

Income Composition and Hukou-Based Rents

The returns to land are distributed to workers who hold the local hukou registration. Village land is collectively owned and its usufruct rights are allocated to registered members; urban land rents accrue to municipal governments and are largely spent on local public goods benefiting registered residents. In our model,

land rents generated in region i are $\Pi_i = R_i^H \overline{\mathcal{H}}_i$. We assume that Π_i is rebated as an equal per-capita lump-sum transfer to all workers registered in i, regardless of where they currently reside. This captures the idea that migrants who leave their rural villages often retain residual claims on village land, even if those claims are attenuated by distance. We abstract from the possibility that prolonged absence might erode these claims; introducing a distance- or duration-dependent attenuation factor would strengthen the home bias and amplify the mechanism we identify. The quantitative effect of such a modification is discussed in Appendix F.

For capital income, as described in Section 3, workers registered in i receive $\overline{r}$ on their capital endowment $\overline{K}_i$, plus an equal per-capita share of aggregate intermediary profits $\Pi^F/\sum_k \overline{L}_k$. In addition, because firms earn operating profits under monopolistic competition, each worker also receives an equal per-capita share of aggregate firm profits, denoted by $\Pi^M/\sum_k \overline{L}_k$, where $\Pi^M \equiv \sum_{j=1}^{N} \sum_{s=1}^{S} \pi_{j,s}$ and $\pi_{j,s}$ is the operating profit of the sector-s firm in region j. The assumption that all firms are owned equally by all workers, regardless of hukou registration or location, ensures that firm ownership does not introduce an additional spatial wedge into migration decisions.

Putting these components together, the nominal income of a worker registered in i employed in region j is

$$I_{i,js} = w_j + \frac{\Pi_i}{\overline{L}_i} + \frac{\overline{r}\overline{K}_i}{\overline{L}_i} + \frac{\Pi^F}{\sum_{k=1}^{N} \overline{L}_k} + \frac{\Pi^M}{\sum_{k=1}^{N} \overline{L}_k}.$$

The first term is labor income earned at the destination. The remaining terms depend only on the workers hukou origin i, not on the destination j. This creates a powerful home bias: even if two regions offer identical wages, prices, and amenities, a worker prefers to remain in or return to their hukou location because migration does not increase non-wage income and, through the hukou friction μ_{ij},

reduces effective real consumption. The associated real income is $y_{i,js} = I_{i,js}P_j^{-\alpha}(R_j^H)^{-(1-\alpha)}$.

C. Producers: Endogenous Technology Choice

This section constitutes the core theoretical contribution of the paper. We depart from the canonical spatial model by allowing firms to optimally select their production function from a menu of available techniques.

The Technology Menu

Our modeling approach builds on the appropriate technology framework of Basu and Weil (1998) and Caselli and Coleman (2006). In those models, a world technology frontier offers a choice of production techniques, and rational producers select the technique that minimizes unit costs given their local factor endowments. We adapt this logic to a within-country spatial setting and introduce a transparent parametric representation of the fundamental trade-off that makes the technology choice both analytically tractable and empirically identifiable.

In each region i and sector s, a representative firm produces a differentiated variety using capital $K_{i,s}$ and labor $L_{i,s}$. The firm has access to a continuum of CES production functions indexed by the substitution parameter ρ, where the admissible set is a compact interval $[\underline{\rho}, \bar{\rho}]$ with $-\infty < \underline{\rho} < 0 < \bar{\rho} < 1$. The production function for a chosen ρ is

$$Y_{i,s} = Z_{i,s}[\gamma(\rho)K_{i,s}^{\rho} + (1-\gamma(\rho))L_{i,s}^{\rho}]^{1/\rho},$$

where $Z_{i,s} > 0$ is an exogenous Hicks-neutral productivity shifter, and $\gamma(\rho) \in (0,1)$ is the capital distribution parameter. The elasticity of substitution between capital and labor is $\sigma = 1/(1-\rho)$. Values of $\rho > 0$ correspond to $\sigma > 1$ (capital and labor are gross substitutes), whereas $\rho < 0$ corresponds to $\sigma < 1$ (gross

complements); the limiting case $\rho \to 0$ recovers the Cobb-Douglas production function.

The crucial assumption that drives our mechanism is that the capital distribution parameter is a strictly increasing, continuously differentiable function of ρ:

$$\gamma'(\rho) > 0.$$

This restriction formalizes a realistic technological trade-off: techniques that permit greater flexibility in substituting between capital and labor are inherently more capital-intensive in their design. Rather than treating this as a primitive assumption, we provide a purely deterministic microfoundation in Appendix C that derives this relationship from a research possibility frontier. In this framework, firms allocate scarce innovation resources between capital-augmenting and labor-augmenting technical improvements. Higher substitutability is technically demanding-it requires the production blueprint to accommodate a wider range of input combinations-and therefore consumes research resources that could otherwise be devoted to raising factor-specific efficiencies. Because a higher substitution elasticity makes the production function more responsive to factor-augmenting technical change in the capital dimension, it increases the marginal benefit of directing research effort toward capital augmentation. When capital is relatively abundant or cheap, firms optimally bias their research toward the capital dimension, making the resulting blueprint more capital-intensive. This deterministic trade-off yields $\gamma'(\rho) > 0$ without any appeal to unmodeled uncertainty, ensuring full consistency with the deterministic general equilibrium in the main text.

One may naturally ask whether the technology menu should be sector-specific. In principle, the research possibility frontier could differ across industries, leading to sector-specific $\gamma_s(\rho)$ functions. Our framework accommodates such sectoral heterogeneity; the TFP variance decomposition remains valid, and sector-specific

menus would likely amplify the endogenous heterogeneity component. We maintain a common $\gamma(\rho)$ in the main exposition for clarity, and we address cross-regional heterogeneity in the menu-the possibility that coastal provinces with superior access to foreign technology face a different frontier than inland provinces-in the quantitative analysis by allowing the intercept of $\gamma(\rho)$ to depend on FDI exposure and trade openness as robustness checks.

Cost Minimization and Optimal Technology Adoption

The firm solves a two-stage problem. In the second stage, taking the technology ρ and factor prices (w_i, r_i) as given, it chooses inputs to minimize unit cost. The conditional factor demand ratio is

$$\frac{K_{i,s}}{L_{i,s}} = \left(\frac{\gamma(\rho)}{1-\gamma(\rho)}\right)^{\frac{1}{1-\rho}} \left(\frac{w_i}{r_i}\right)^{\frac{1}{1-\rho}}.$$

The minimized unit cost function is $c(w_i, r_i; \rho) = \frac{r_i}{Z_{i,s}} \mathcal{C}(\omega_i; \rho)$, where $\omega_i \equiv w_i/r_i$ is the wage-rental ratio and

$$\mathcal{C}(\omega_i; \rho) \equiv [\gamma(\rho)^{\frac{1}{1-\rho}} + (1-\gamma(\rho))^{\frac{1}{1-\rho}} \omega_i^{-\frac{\rho}{1-\rho}}]^{-\frac{1-\rho}{\rho}}.$$

In the first stage, the firm selects $\rho \in [\underline{\rho}, \bar{\rho}]$ to minimize $\mathcal{C}(\omega_i; \rho)$. The first-order necessary condition is $\partial \ln \mathcal{C}(\omega_i; \rho)/\partial \rho = 0$. Defining the auxiliary function $\Phi(\rho, \omega_i) \equiv \gamma(\rho)^{\frac{1}{1-\rho}} + (1-\gamma(\rho))^{\frac{1}{1-\rho}} \omega_i^{-\frac{\rho}{1-\rho}}$, we have $\ln \mathcal{C} = -\frac{1-\rho}{\rho} \ln \Phi$, and the first-order condition becomes

$$\frac{1}{\rho^2} \ln \Phi - \frac{1-\rho}{\rho} \frac{\Phi_\rho}{\Phi} = 0,$$

where $\Phi_\rho \equiv \partial \Phi / \partial \rho$. A unique interior minimum requires the second-order condition $\partial^2 \ln \mathcal{C}/\partial \rho^2 > 0$.

Lemma 1 (Comparative Statics of Technology Choice). Assume $\gamma'(\rho) > 0$ for all $\rho \in [\underline{\rho}, \bar{\rho}]$, and that the second-order condition for a unique interior minimum holds at ρ_i^*. Then $d\rho_i^*/d\omega_i < 0$.

The proof, provided in Appendix A, establishes that the mixed partial derivative $\partial^2 \ln \mathcal{C}/\partial\rho\partial\omega$ is strictly positive under the maintained assumptions. A sufficient condition is that the elasticity of γ with respect to ρ exceeds a threshold that depends on ρ and ω. This condition has a clear economic interpretation. The positive association between the substitution elasticity and capital intensity must be strong enough that firms find it worthwhile to adopt a more flexible, capital-intensive technology when labor becomes more expensive. If $\gamma'(\rho)$ were too small-if more flexible technologies were only marginally more capital-intensive-firms might prefer to stay at a corner solution with the lowest feasible ρ, simply substituting capital for labor along a fixed isoquant rather than switching to a more flexible production process. The threshold reflects the trade-off: a higher ρ reduces unit costs through two channels-it raises the substitution elasticity (making it easier to economize on expensive labor) and it increases capital intensity (further reducing costs when capital is cheap)-but the second channel only operates if $\gamma'(\rho)$ is large enough. Under the empirically relevant parameterizations of $\gamma(\rho)$, including the linear specification $\gamma(\rho) = \theta_0 + \theta_1\rho$ with our estimated $\hat{\theta}_1 = 0.34$, the condition is easily satisfied for all Chinese provinces in our sample.

Let $c_i^* \equiv \mathcal{C}(\omega_i; \rho^*(\omega_i))$ denote the minimized unit cost component. The full unit cost for a firm in region i and sector s is $c_{i,s}^* = r_i c_i^*/Z_{i,s}$. Unit costs are lower in regions with higher fundamental productivity, lower capital rental rates, and wage-rental ratios that are more favorable given the chosen technology. The spatial variation in c_i^* induced by differences in ω_i -and the corresponding

differences in ρ_i^* -is the source of the endogenous TFP heterogeneity that we analyze below.

D. Goods Markets and Monopolistic Competition

We assume that varieties produced in different regions within the same sector are imperfect substitutes. Sectoral composite consumption in region j is a CES aggregate over varieties sourced from all regions with trade elasticity $\sigma_s > 1$ and iceberg trade costs $\tau_{ij,s} \geq 1$, with $\tau_{ii,s} = 1$. The sectoral price index is

$$P_{j,s} = [\sum_{i=1}^{N} (\tau_{ij,s} p_{i,s})^{1-\sigma_s}]^{\frac{1}{1-\sigma_s}},$$

where $p_{i,s}$ denotes the mill price. Cost minimization by competitive distributors yields the demand schedule $q_{ij,s} = (\tau_{ij,s} p_{i,s}/P_{j,s})^{-\sigma_s} C_{j,s}$ and the gravity equation for trade shares $\pi_{ij,s} \equiv (\tau_{ij,s} p_{i,s})^{1-\sigma_s}/\sum_k (\tau_{kj,s} p_{k,s})^{1-\sigma_s}$.

The choice of market structure is of first-order importance for the mechanism we study, and we therefore depart from the perfectly competitive benchmark adopted in many quantitative spatial models. Under perfect competition with homogeneous goods, the law of one price would force all but the lowest-cost region out of the market, eliminating the very cross-regional variation in production techniques that is our object of study. Even under Armington differentiation with perfectly competitive pricing, the absence of a meaningful demand-side channel would sever the link between local factor costs and equilibrium output scales that is necessary for heterogeneous technologies to be observed in equilibrium. We adopt instead a monopolistically competitive market structure a la Dixit and Stiglitz (1977), in which each region produces a distinct variety and the representative firm in each region-sector chooses its mill price to

maximize operating profits, taking the sectoral price index $P_{j,s}$ as given. The profit maximization problem yields the standard constant-markup pricing rule

$$p_{i,s} = \frac{\sigma_s}{\sigma_s - 1} c^*_{i,s} = \frac{\sigma_s}{\sigma_s - 1} \frac{r_i c^*_i}{Z_{i,s}}.$$

The markup $\mu_s \equiv \sigma_s/(\sigma_s - 1) > 1$ is common to all firms within the same sector, regardless of their location. This property is crucial for our analysis. Because the markup is constant within a sector, it does not introduce any cross-regional variation in revenue productivity beyond what is already captured by differences in unit costs $c^*_{i,s}$ and capital rental rates r_i. In particular, the firms optimal technology choice ρ^*_i -which depends only on the wage-rental ratio ω_i -is completely unaffected by the markup, since μ_s factors out multiplicatively from the profit-maximization problem. This clean separation between the technology choice margin (driven by factor prices) and the pricing margin (driven by product market power) is what allows us to isolate the endogenous technology heterogeneity channel from the confounding effects of variable markups.

The monopolistic competition structure also plays an essential role in our empirical identification strategy. In the data, observed capital shares reflect not only technological parameters but also the pricing decisions of firms with market power. The control function approach of De Loecker and Warzynski (2012), which we employ to purge factor shares of markup variation, is fundamentally grounded in the premise that firms operate in an imperfectly competitive environment. By making monopolistic competition an integral part of our theoretical framework-rather than treating markup correction as an ex-post robustness check-we ensure full coherence between the models structure and its empirical implementation.

Operating profits earned by the sector-s firm in region i are

$$\pi_{i,s} = (p_{i,s} - c^*_{i,s}) Y_{i,s} = \frac{1}{\sigma_s - 1} c^*_{i,s} Y_{i,s}.$$

These profits are assumed to be distributed as equal per-capita dividends to all workers in the economy, as described in Section 4.3. Aggregate firm profits are $\Pi^M \equiv \sum_{j=1}^{N} \sum_{s=1}^{S} \pi_{j,s}$.

Final consumption expenditure on sector s in region j is $E_{j,s} = \alpha_s \alpha I_j L_j$. Total revenue of the sector-s firm in region i equals the sum of sales to all destinations, $R_{i,s} = \sum_{j=1}^{N} \pi_{ij,s} E_{j,s}$, and operating profits are $\pi_{i,s} = R_{i,s}/\sigma_s$.

E. General Equilibrium and Constrained Efficiency

A competitive equilibrium consists of factor prices $\{w_i, r_i, R_i^H\}$, goods prices $\{P_{i,s}, p_{i,s}\}$, technology choices $\{\rho_i^*\}$, quantities $\{L_{i,s}, K_{i,s}, \pi_{ij,s}, C_{i,s}, H_i\}$, and incomes and migration shares $\{I_i, m_{i,js}\}$ such that migration is optimal, technology minimizes unit cost, firms set profit-maximizing prices according to the constant-markup rule, factor and goods markets clear, and the income identity

$$I_j L_j = w_j L_j + \sum_{i=1}^{N} \sum_{s=1}^{S} m_{i,js} \overline{L}_i \cdot \left(\frac{\Pi_i}{\overline{L}_i} + \frac{\overline{r}\overline{K}_i}{\overline{L}_i} + \frac{\Pi^F}{\sum_k \overline{L}_k} + \frac{\Pi^M}{\sum_k \overline{L}_k} \right)$$

holds for each region j. Existence is established via Kakutanis fixed-point theorem (Appendix B).

To provide a rigorous normative foundation, we formulate a constrained planners problem. The planner maximizes a utilitarian social welfare function subject to the same hukou costs μ_{ij} and capital-market wedges τ_i^K that constrain private agents. Crucially, the capital wedge represents a real resource cost: to provide one unit of productive capital to firms in region i, the planner must divert τ_i^K units of capital from the aggregate endowment. The planner can redistribute the rents from financial intermediation and the operating profits from monopolistic competition, but cannot eliminate either the capital wedge or the product differentiation that gives rise to markups. Proposition 3 establishes that

the decentralized equilibrium coincides with the planners solution, so the endogenous technology heterogeneity is constrained-efficient. The proof, in Appendix E, verifies that the planners first-order conditions for technology choice, factor allocation, migration, and goods markets are identical to those of the decentralized equilibrium. The presence of constant within-sector markups does not create a wedge between the two allocations, because the planner faces the same CES demand structure and therefore the same markup structure as private firms.

This result requires careful interpretation. The frictions μ_{ij} and τ_i^K are the sources of inefficiency relative to the first-best. The planners problem does not deny this; it simply respects the same frictions as the market. The point is that given these frictions, the market achieves the second-best allocation, and the technology dispersion that emerges is part of that optimum. We reserve the term "misallocation" for deviations from this second-best benchmark that could be eliminated without removing the deep spatial frictions-for instance, firm-specific subsidies or politically motivated credit allocation not captured by region-level wedges. The technology choice component $Var(\ln T(\omega_i))$ is not misallocation by this definition, because it cannot be reduced without either removing the frictions or mandating technologies that would increase unit costs. We demonstrate this concretely in Section 9: a counterfactual that forces all regions to adopt a common technology while keeping the frictions in place reduces aggregate output. The constant markups arising from monopolistic competition do not themselves constitute misallocation; they are simply a feature of product differentiation under CES preferences. However, deviations of firm-level markups from the sectoral mean-which the De Loecker and Warzynski (2012) methodology is designed to detect-would represent genuine misallocation in our framework.

III. TFP Variance Decomposition: Separating Endogenous Technology Choice from Misallocation

The theoretical framework developed in Sections 2-7 yields a central analytical result: in the presence of persistent spatial frictions, measured TFP dispersion across regions and firms is a composite of several fundamentally distinct components. This section derives the exact decomposition formulas, discusses their economic interpretation, and-crucially-delineates which components reflect efficient adaptation to constraints and which represent genuine misallocation. The decomposition serves two purposes. First, it provides the analytical foundation for the quantitative variance decomposition and counterfactual simulations in Section 9. Second, it offers a conceptual framework that clarifies the relationship between our mechanism and the standard misallocation accounting of Hsieh and Klenow (2009, henceforth HK).

The core insight is straightforward to state but has far-reaching implications. When firms in different regions face persistently different factor prices-here, because of the *hukou* system and capital-market segmentation-they optimally select different technologies from the menu described in Section 5. These different technologies imply different unit cost functions, and hence different levels of measured TFP, even if all firms operate with identical fundamental productivity Z and face no firm-specific distortions whatsoever. The resulting dispersion in measured TFP is *not* misallocation in any economically meaningful sense: it is the constrained-optimal outcome given the spatial frictions, and it would vanish automatically if the frictions were removed. Traditional misallocation accounting, which treats all TFP dispersion as evidence of removable wedges, therefore conflates two distinct phenomena. Our decomposition provides a rigorous method for separating them.

A. Defining Measured TFP: TFPQ and TFPR

A necessary preliminary is to distinguish carefully between two measures of total factor productivity that appear in the literature.

Quantity-based TFP (TFPQ). In the theoretical model of Section 5, the quantity-based TFP of the representative firm in region i and sector s is defined directly from the physical production function. Using the optimal technology $\rho_i^* \equiv \rho^*(\omega_i)$ and the minimized unit cost function, we write output as

$$Y_{i,s} = Z_{i,s}\left[\gamma(\rho_i^*)K_{i,s}^{\rho_i^*} + (1-\gamma(\rho_i^*))L_{i,s}^{\rho_i^*}\right]^{1/\rho_i^*},$$

where $Z_{i,s}$ is exogenous Hicks-neutral productivity. The minimized unit cost component, derived in Section 5.2, is $c_i^* \equiv \mathcal{C}(\omega_i; \rho_i^*)$, with $\omega_i \equiv w_i/r_i$. We define TFPQ as

$$A_{i,s} \equiv \frac{Y_{i,s}}{K_{i,s}^{\gamma(\rho_i^*)} L_{i,s}^{1-\gamma(\rho_i^*)}} = \frac{Z_{i,s}}{c_i^*}.$$

Taking logs,

$$\ln A_{i,s} = \ln Z_{i,s} - \ln T(\omega_i),$$

where we introduce the compact notation $T(\omega_i) \equiv c_i^* = \mathcal{C}(\omega_i; \rho^*(\omega_i))$ for the minimized unit cost component expressed as a function of the wage-rental ratio alone. The function $T(\omega_i)$ is strictly decreasing in ω_i under the conditions of Lemma 1, meaning that regions with higher wage-rental ratios (labor more expensive relative to capital) exhibit lower unit costs after optimal technology choice.

Revenue-based TFP (TFPR). HKs empirical analysis uses revenue-based TFP, which incorporates output prices. Under the monopolistic competition structure of Section 6, the firms product price is $P_{i,s} = \mu_s \cdot c_{i,s}^*$, where $c_{i,s}^* = r_i c_i^*/Z_{i,s}$ is the full unit cost, r_i is the capital rental rate, and $\mu_s \equiv \sigma_s/(\sigma_s - 1)$ is the sector-specific

constant markup. Revenue TFP for the representative firm in region i and sector s is defined as the Solow residual from a revenue production function:

$$\text{TFPR}_{i,s} \equiv \frac{P_{i,s}Y_{i,s}}{\tilde{K}_{i,s}^{\tilde{\gamma}} L_{i,s}^{1-\tilde{\gamma}}},$$

where $\tilde{\gamma}$ is a benchmark capital elasticity used for cross-sectional comparability. In the standard HK framework, $\tilde{\gamma}$ is set to the industry average cost share of capital; in our analysis, we use either the national average $\bar{\gamma}$ or, in robustness checks, the region-specific $\gamma(\rho_i^*)$. Substituting the expressions for $P_{i,s}$, $Y_{i,s}$, and the factor demands yields, after algebraic simplification,

$$\ln\text{TFPR}_{i,s} = \underbrace{\ln\mu_s}_{\text{industry constant}} + \ln Z_{i,s} - \ln T(\omega_i) + (\tilde{\gamma} - \gamma(\rho_i^*))\ln\left(\frac{K_{i,s}}{L_{i,s}}\right) + \text{constant}.$$

The key observation is that lnTFPR contains two sources of variation beyond exogenous productivity $Z_{i,s}$: the endogenous technology choice term $\ln T(\omega_i)$, and a capital-intensity adjustment that arises whenever the benchmark elasticity $\tilde{\gamma}$ differs from the local $\gamma(\rho_i^*)$. Under the constant-markup assumption, the industry term $\ln\mu_s$ is common to all firms within a sector and does not contribute to within-sector TFPR dispersion.

This dual definition is essential for meaningful dialogue with HK. HKs central object is the cross-sectional dispersion of lnTFPR across firms within narrowly defined industries. They interpret all such dispersion-after accounting for measurement error-as evidence of firm-specific distortionary wedges. Our decomposition reveals that even at the representative-firm level (i.e., before introducing firm-specific heterogeneity), lnTFPR already exhibits systematic cross-regional dispersion driven by $\ln T(\omega_i)$. In Section 9, we extend the analysis to firm-level data and show that this regional component accounts for a substantial fraction of the total within-industry TFPR variance that HK attribute entirely to misallocation.

B. The Frictionless Benchmark: Proposition 1

We first establish the case in which no misallocation or endogenous dispersion exists.

Proposition 1 (Frictionless Benchmark). If $\mu_{ij} = 1$ for all i, j and $\tau_i^K = 1$ for all i, then in equilibrium: (i) factor prices equalize across regions, $w_i = w$ and $r_i = r$ for all i, implying a common wage-rental ratio $\omega_i = \omega$; (ii) all regions adopt the same technology $\rho_i^* = \rho^* = \operatorname{argmin}_{\rho} \mathcal{C}(\omega; \rho)$; and (iii) the cross-sectional variance of log TFPQ reduces to the variance of exogenous productivity alone:

$$Var(\ln A_{i,s}) = Var(\ln Z_{i,s}).$$

For TFPR, using a common benchmark elasticity $\tilde{\gamma}$, we additionally have $Var(\ln \text{TFPR}_{i,s}) = Var(\ln Z_{i,s})$.

Proof. With no migration costs and no capital wedges, labor and capital flow freely across regions until factor prices equalize. Equation (3) implies $w_i = w$ and $r_i = r$ for all i, hence $\omega_i = \omega$. By Lemma 1, the unique cost-minimizing ρ^* is identical for all regions. The unit cost component $T(\omega_i) = T(\omega)$ is a common constant and drops out of the cross-sectional variance. For TFPR, factor price equalization implies identical capital-labor ratios (up to industry composition), and the adjustment term in (8.3) vanishes when $\tilde{\gamma} = \gamma(\rho^*)$. ▫

Proposition 1 formalizes the standard implicit assumption in the misallocation literature: absent distortions, TFP dispersion reduces to fundamental productivity differences. It also clarifies the necessary conditions for this result-not merely the absence of firm-level wedges, but also perfect factor mobility across space. In an economy like China, where institutional barriers to migration and capital flows are first-order features of the economic landscape, Proposition 1 provides an instructive but counterfactual benchmark.

C. Decomposition under Spatial Frictions: Proposition 2

We now turn to the empirically relevant case with binding spatial frictions.

Proposition 2 (Decomposition under Spatial Frictions). Under the spatial frictions described in Section 3 ($\mu_{ij} \geq 1$, $\tau_i^K \geq 1$), the cross-sectional variance of log TFPQ across region-sector cells is decomposed as:

$$Var(\ln A_{i,s}) = Var(\ln Z_{i,s}) + Var(\ln T(\omega_i)) - 2Cov(\ln Z_{i,s}, \ln T(\omega_i)).$$

For log TFPR within a given sector s, the decomposition is:

$$\begin{aligned} Var(\ln \text{TFPR}_{i,s}) = \quad & Var(\ln Z_{i,s}) + Var(\ln T(\omega_i)) + Var(\ln r_i) \\ & -2Cov(\ln Z_{i,s}, \ln T(\omega_i)) \\ & -2Cov(\ln Z_{i,s}, \ln r_i) + 2Cov(\ln T(\omega_i), \ln r_i) \\ & + \text{ terms involving } (\tilde{\gamma} - \gamma(\rho_i^*)). \end{aligned}$$

Proof. Equation (8.4) follows directly from (8.2) and the definition of variance. For TFPR, taking logs of (8.3) and computing the cross-sectional variance within sector s (so that $\ln \mu_s$ is constant) yields the expression, with the covariance terms arising from the inclusion of $\ln r_i$ and the capital-intensity adjustment. The sign of the covariance terms depends on the spatial equilibrium: in China, productive coastal provinces tend to have lower τ_i^K and hence lower r_i, creating a negative covariance between $\ln Z$ and $\ln r_i$ that amplifies TFPR dispersion relative to TFPQ dispersion. ▫

The decomposition has a clear economic interpretation, which we now elaborate.

$Var(\ln Z_{i,s})$ **- Exogenous productivity dispersion.** This component reflects fundamental differences in technological know-how, climate, infrastructure quality, or other factors that shift the production possibilities frontier independently of factor prices. It is the only component that would remain in a frictionless world (Proposition 1). To the extent that such productivity differences

reflect persistent geographic or institutional advantages, they are a feature of the economic environment, not a distortion.

$Var(\ln T(\omega_i))$ **- Endogenous technology choice dispersion.** This is the novel component introduced by our framework. It captures the fact that, when ω_i differs across regions due to spatial frictions, firms optimally choose different points on the technology menu $\gamma(\rho)$. Regions with high ω_i (labor expensive relative to capital) choose higher ρ_i^*, implying more flexible, capital-intensive technologies and-through the unit cost function $T(\omega_i)$-systematically different measured TFP. By Proposition 3, this dispersion is constrained-efficient: no policy that respects the underlying frictions can reduce it without increasing aggregate production costs. The counterfactual simulation in Section 9, in which forcing uniform technology reduces aggregate output by 3.8%, provides direct quantitative evidence for this claim.

$Cov(\ln Z_{i,s}, \ln T(\omega_i))$ **- Covariance between productivity and technology choice.** If regions with high exogenous productivity also have high wage-rental ratios-as is plausible in China, where productive coastal regions face tight labor markets due to *hukou* restrictions-this covariance will be negative (since $T(\omega)$ is decreasing in ω). A negative covariance means that the two components partially offset each other, reducing total TFP dispersion. This offset has an important economic interpretation: technological adaptation partially compensates for underlying productivity differences by allowing high-productivity regions to adopt technologies that economize on their scarce factor.

$Var(\ln r_i)$ **- Capital price dispersion.** This term appears only in the TFPR decomposition and reflects the fact that capital-market wedges τ_i^K directly affect measured revenue productivity through the rental rate $r_i = \tau_i^K \bar{r}$. Regions with preferential access to credit (low τ_i^K) enjoy a lower r_i, which mechanically raises measured TFPR by reducing the denominator of the unit cost expression. This

component is a direct consequence of the capital-market segmentation described in Section 3.

Covariance terms involving $\ln r_i$**.** The sign and magnitude of these terms depend on the spatial equilibrium. In China, coastal provinces tend to have both high $Z_{i,s}$ and low r_i (high productivity and cheap capital), generating a negative $Cov(\ln Z_{i,s}, \ln r_i)$ that amplifies TFPR dispersion. This amplification is a distinctive prediction of our model that distinguishes the TFPQ and TFPR decompositions: TFPR dispersion can be substantially larger than TFPQ dispersion due to the compounding effects of capital-market segmentation on both the technology choice margin (through ω_i) and the direct cost channel (through r_i).

The distinction between TFPQ and TFPR matters for the dialogue with HK. HK measure TFPR and treat its dispersion as a sufficient statistic for misallocation. Our decomposition shows that TFPR dispersion is, from first principles, a convolution of exogenous productivity, endogenous technology choice, and spatial variation in factor prices. Using TFPR dispersion as a direct measure of misallocation therefore requires the strong assumption that $T(\omega_i)$ and r_i are constant across the comparison set. For between-region comparisons within the same industry in China, this assumption is demonstrably false.

D. From Regions to Firms: Nesting HKs Within-Industry Misallocation

Proposition 2 operates at the region-sector level, decomposing the between-group variation in TFP. However, HKs analysis, and the broader misallocation literature, focuses on firm-level dispersion within industries. To bridge this gap, we now introduce firm-level heterogeneity.

Following HK, suppose that within each region i and sector s, there is a continuum of firms indexed by f. Each firm draws an idiosyncratic productivity

Z_f and faces firm-specific distortions to capital and labor, denoted τ_f^K and τ_f^L, which scale the effective factor prices to $(1+\tau_f^K)r_i$ and $(1+\tau_f^L)w_i$. Crucially, all firms in region i face the same local factor prices (w_i, r_i) and have access to the same optimal technology ρ_i^*-the technology choice is made at the regional level, reflecting the long-run nature of production function selection compared to the potentially transient nature of firm-specific wedges. The firms production technology is thus

$$Y_f = Z_f\left[\gamma(\rho_i^*)K_f^{\rho_i^*} + (1-\gamma(\rho_i^*))L_f^{\rho_i^*}\right]^{1/\rho_i^*}.$$

The firms optimal capital-labor ratio satisfies the first-order condition (12), and its log TFPR, following the derivation in Appendix G, can be written as

$$\ln\text{TFPR}_f = \ln\Xi(\omega_i;\rho_i^*) + \ln Z_f + F(\tau_f^K, \tau_f^L; \rho_i^*),$$

where $\ln\Xi(\omega_i;\rho_i^*) \equiv \ln\mu_s - \ln T(\omega_i) + \tilde{\gamma}\ln r_i + (1-\tilde{\gamma})\ln w_i$ collects all terms that vary only at the regional level, and $F(\cdot)$ is a function capturing the effect of firm-specific distortions on measured TFPR.

Equation (8.7) is the key structural relationship. It shows that firm-level TFPR is the sum of three components:

- **A regional technology component** $\ln\Xi(\omega_i;\rho_i^*)$, which is common to all firms in the same province and reflects the constrained-optimal response to local factor prices;
- **An idiosyncratic productivity component** $\ln Z_f$, capturing genuine heterogeneity in firm capabilities;
- **A distortion component** $F(\tau_f^K, \tau_f^L; \rho_i^*)$, which captures the extent to which firm-specific wedges drive a gap between the firms actual TFPR and the level it would achieve under efficient factor allocation within the region.

The variance of $\ln\text{TFPR}_f$ across all firms can now be decomposed as

$$Var(\ln \text{TFPR}_f) = Var(\ln \bar{\text{TFPR}}_{i,s}) + \mathbb{E}\left[Var(\ln \text{TFPR}_f | i, s)\right],$$

where $\ln \bar{\text{TFPR}}_{i,s}$ is the average log TFPR in province i and sector s. The between-group variance $Var(\ln \bar{\text{TFPR}}_{i,s})$ is precisely what Proposition 2 analyzes at the representative-firm level. The within-group variance $\mathbb{E}[Var(\ln \text{TFPR}_f | i, s)]$ captures the firm-specific dispersion that is the focus of HK.

Moreover, the between-group variance itself can be further separated using the structural function $\ln \Xi$:

$$Var(\ln \bar{\text{TFPR}}_{i,s}) = Var(\ln \Xi(\omega_i; \rho_i^*)) + \text{residual regional misallocation},$$

where $Var(\ln \Xi(\omega_i; \rho_i^*))$ is the endogenous technology choice contribution, and the residual captures any systematic regional-level factors (such as province-specific taxes, subsidies, or bureaucratic quality) not already embodied in μ_{ij} and τ_i^K.

This nested decomposition is implemented empirically in Section 9, where we show that the endogenous technology choice component accounts for 26.8% of the total firm-level TFPR variance in Chinese manufacturing. The within-group, firm-specific distortion component accounts for about 43.7%. The remainder consists of exogenous productivity heterogeneity and residual noise.

E. Relationship to the HK Framework

Our decomposition stands in a complementary but critical relationship to HK.

HKs central exercise is to compute the potential gains from equalizing marginal revenue products across firms within narrowly defined industries in China and India. Their calculation assumes that all firms share an identical production function, characterized by industry-specific factor elasticities. Under this assumption, any dispersion in measured TFPR across firms reflects distortionary

wedges, and eliminating those wedges would raise manufacturing TFP by 30-50% in China.

Our framework enriches this picture in two ways.

First, it identifies a source of TFPR dispersion-the endogenous technology choice component $Var(\ln\Xi)$ -that is *not* misallocation but rather the efficient response to spatial frictions. This component would be incorrectly counted as removable distortion in the HK calculation. Our counterfactual analysis in Section 9 shows that, after removing the firm-specific distortions within each region while respecting the spatial frictions, the potential TFP gain is only +24.9%, substantially below the +43.2% obtained from replicating the HK calculation on our data. The 18.3 percentage point difference is exactly the overestimation due to ignoring endogenous technology choice.

Second, our framework clarifies that the path to achieving even larger gains-+51.6% in our Scenario C-requires removing the deep spatial frictions (the *hukou* system and capital-market segmentation) that cause factor prices, and hence technologies, to differ across regions in the first place. This is a fundamentally different policy agenda from the firm-level reallocation emphasized by HK. It implies that partial reforms targeting within-industry factor allocation, while beneficial, will capture only a fraction of the potential efficiency gains if the spatial institutional barriers remain in place.

Our analysis does not, however, contradict HKs fundamental insight. Within-group dispersion-the firm-specific distortion component in our decomposition-remains large and economically significant. The point is simply that not all TFPR dispersion is created equal: some of it is endogenous adaptation to constraints, not a deviation from efficiency given those constraints. Distinguishing between these sources is essential for both accurate measurement and sound policy design.

F. Constrained Efficiency and Normative Implications

The theoretical status of the decomposition is clarified by Proposition 3 (Constrained Efficiency, proved in Appendix E). The constrained planner, who must respect the same *hukou* costs μ_{ij} and capital-market wedges τ_i^K as private agents, would choose exactly the same allocation-including the same region-specific technologies ρ_i^* -as the decentralized equilibrium. This implies that $Var(\ln T(\omega_i))$ is not just an empirical artifact; it is a feature of the constrained-optimal allocation. Any policy that reduced this variance without relaxing the underlying frictions-for example, by subsidizing capital in labor-abundant regions to equalize capital-labor ratios-would reduce aggregate welfare by distorting the technology choice margin.

The normative implication is clear: the appropriate target of welfare-improving reform is the spatial frictions themselves, not the measured TFP dispersion they generate. This conclusion parallels the classic theory of the second-best (Lipsey and Lancaster, 1956): in the presence of irreducible distortions, the optimal response is not to eliminate all observable inequalities but to optimize subject to the constraints. Our quantitative contribution is to show that this theoretical point has first-order empirical consequences for the misallocation literatures headline numbers.

G. Taking Stock: From Decomposition to Quantification

The analytical results of this section provide the scaffolding for the quantitative analysis in Section 9. The empirical strategy can be summarized in three steps. First, we estimate the technology menu $\gamma(\rho)$ and identify the optimal ρ_i^* for each Chinese province using the structural equations derived from Lemma 1. Second, we calibrate the spatial frictions μ_{ij} and τ_i^K from observed migration flows and capital costs. Third, using firm-level data, we implement the nested variance decomposition of equations (8.8)-(8.9) and conduct counterfactual simulations

that correspond to alternative policy scenarios. The results, as previewed above, provide a quantitative separation of endogenous technology choice from genuine misallocation and yield corrected estimates of the potential gains from reforms of different scope.

IV. Parameter Calibration, Empirical Strategy, and Quantitative Analysis

This section translates the theoretical framework into estimable structural parameters and testable empirical decompositions, and engages directly with the canonical misallocation accounting of Hsieh and Klenow (2009, hereafter HK) at the firm level. The empirical design follows a core logic: if the mechanism "factor endowments → technology choice → productivity dispersion" revealed by the theory is genuine, then HK's practice of attributing all TFPR dispersion to distortionary wedges will produce systematic overestimation. We sequentially calibrate the technology menu, the spatial friction parameters, and firm-level distortions. Relying on the theoretically derived structural form of TFPR, we decompose firm-level productivity variance into three components: endogenous technology choice, residual regional misallocation, and firm-specific distortions. Finally, through multiple counterfactual simulations we quantify the corrected potential gains in manufacturing TFP.

A. Data Sources and Construction of Key Variables

1. Data

We use the 1998-2013 Annual Survey of Industrial Firms (ASIF) of China, which covers all state-owned industrial enterprises and non-state-owned enterprises with annual sales above 5 million yuan. Data cleaning follows the standard procedures of Brandt, Van Biesebroeck, Wang, and Zhang (2017): observations with missing or negative values for value added, employment, net

fixed assets, or intermediate inputs are dropped; firms with fewer than 8 employees are excluded to remove non-formal units; nominal variables are deflated using industry-specific output and input price deflators; real capital stocks are constructed via the perpetual inventory method with a depreciation rate of 9%; to mitigate the influence of outliers, the main financial ratios and TFPR are winsorized at the 1st and 99th percentiles within each 2-digit industry-year. The final estimation sample comprises approximately 28 provinces, 29 manufacturing industries, and around 2.3 million firm-year observations.

Province-level auxiliary data come from the *China Statistical Yearbook* and the benchmark lending rates published by the Peoples Bank of China; these are used to construct the provincial user cost of capital. Prior information on bilateral interprovincial migration flows and the intensity of *hukou* barriers is taken from Tombe and Zhu (2019) and Poncet (2006).

2. Construction of Factor Prices and Factor Shares

Wage rate. The firm-level average wage is obtained by dividing the total wage bill by employment; province-industry-level average wages, weighted by total revenue, are taken as the measure of the regional wage w_i. This treatment implicitly assumes that labor is imperfectly mobile across provinces but highly mobile across industries within a province, consistent with the regional labor market setup of the theoretical model (Sections 2 and 3).

User cost of capital. The regional user cost of capital is computed according to the standard Hall-Jorgenson formula: $r_i = (R_i + \delta)p_{K,i}$, where R_i is the provincial one-year benchmark lending rate published by the People's Bank of China (annual average), $\delta = 0.09$ is the economic depreciation rate, and $p_{K,i}$ is the provincial investment goods price deflator (base year 1998). Here r_i corresponds directly to the rental price of capital faced by firms in the theoretical model (Section 3). In

the presence of provincial capital wedges τ_i^K, its relationship to the national benchmark rate $\bar{r}$ is $r_i = \tau_i^K \bar{r}$.

Wage-rental ratio. The ratio of regional factor prices is defined as $\omega_i \equiv w_i/r_i$. This is the core state variable driving the technology choice mechanism (see Lemma 1, and equations (7)-(8) in the text).

Markup correction and true factor elasticities. Under monopolistic competition, the first-order condition for cost minimization implies that the revenue share of a factor equals its output elasticity divided by the firm's markup. Hence, using revenue shares directly would systematically underestimate the true output elasticities. We apply the control function approach of De Loecker and Warzynski (2012, DLW) that exploits the flexible nature of intermediate inputs to estimate firm-level markups μ_{ft} and recover the true elasticities. This procedure is fully consistent with the constant industry markup assumed in the theoretical Section 6, allowing only for transient firm-level heterogeneity:

- For each 2-digit manufacturing industry, a translog gross production function is estimated:

$$y_{ft} = \beta_k k_{ft} + \beta_l l_{ft} + \beta_m m_{ft} + \frac{1}{2}\beta_{kk} k_{ft}^2 + \frac{1}{2}\beta_{ll} l_{ft}^2 + \frac{1}{2}\beta_{mm} m_{ft}^2 + \beta_{kl} k_{ft} l_{ft} + \beta_{km} k_{ft} m_{ft} + \beta_{lm} l_{ft} m_{ft} + \omega_{ft} + \varepsilon_{ft},$$

where lower-case letters are in logs, ω_{ft} is a productivity shock observable to the firm but not to the econometrician, and ε_{ft} is i.i.d. measurement error. Identification employs the two-stage procedure of Ackerberg, Caves, and Frazer (2015), using intermediate inputs as the proxy variable under the timing assumption that labor is chosen before intermediate inputs.

- The output elasticity of materials is $\hat{\theta}_{ft}^M = \partial y_{ft}/\partial m_{ft}$. Combined with the observed revenue share of materials s_{ft}^M, the markup is estimated as:

$$\hat{\mu}_{ft} = \frac{\hat{\theta}_{ft}^{M}}{s_{ft}^{M}}.$$

- The true output elasticities of capital and labor are recovered as $\hat{\theta}_{ft}^{K} = s_{ft}^{K} \cdot \hat{\mu}_{ft}$ and $\hat{\theta}_{ft}^{L} = s_{ft}^{L} \cdot \hat{\mu}_{ft}$, respectively.
- To link with the representative firm setup of the theoretical model (Section 5), we aggregate the firm-level true capital elasticities to the province-industry level by taking revenue-weighted averages, yielding $\hat{\gamma}_{is}$. These will be treated as the empirical counterparts of the theoretical capital distribution parameter $\gamma(\rho)$.

B. Identification and Estimation of the Technology Menu

1. Identifying Equations

The theoretical model shows that, given the regional wage-rental ratio ω_i, the representative firm chooses the substitution elasticity ρ to minimize the unit cost $\mathcal{C}(\omega_i; \rho)$. The first-order necessary condition is (see equation (8) in the main text):

$$\frac{\partial \ln \mathcal{C}(\omega_i; \rho)}{\partial \rho} = 0.$$

At the same time, from the factor demands of the CES production function the true capital output elasticity is given by (see the factor-share derivation, equation (9)):

$$\gamma_{\text{theory}}(\rho, \omega_i) = \frac{\gamma(\rho)^{\frac{1}{1-\rho}}}{\gamma(\rho)^{\frac{1}{1-\rho}} + (1-\gamma(\rho))^{\frac{1}{1-\rho}} \omega_i^{\frac{\rho}{1-\rho}}}.$$

We assume a linear technology menu $\gamma(\rho) = \theta_0 + \theta_1 \rho$. For every province-industry observation unit $g = (i, s)$, we take the markup-corrected capital elasticity $\hat{\gamma}_g$ as the measurement of γ_{theory}, and together with the computed ω_i, we solve the system of equations (8) and (9). Given a parameter vector θ, this yields

an implicit optimal substitution elasticity $\rho_g(\theta)$ and a model-implied capital share $s_K(\rho_g(\theta), \omega_i; \theta) \equiv \gamma_{\text{theory}}(\rho_g(\theta), \omega_i)$.

2. Estimation and Results

The parameter vector $\theta = (\theta_0, \theta_1)$ is estimated by nonlinear least squares, minimizing the objective function:

$$\min_{\theta_0,\theta_1} \sum_g [\hat{\gamma}_g - s_K(\rho_g(\theta), \omega_i; \theta)]^2.$$

Here $\rho_g(\theta)$ is the numerical solution that satisfies the first-order condition (8) for the given θ. Standard errors are clustered at the province level to account for within-province correlation. Appendix H provides the algorithmic details.

The estimation yields:

$$\hat{\theta}_0 = 0.28 \ (0.04), \qquad \hat{\theta}_1 = 0.29 \ (0.06),$$

where the parentheses contain province-clustered standard errors; both parameters are significant at the 1% level. The result $\hat{\theta}_1 > 0$ strongly supports the core hypothesis of the theory that $\gamma'(\rho) > 0$ (Section 5.1): a higher elasticity of factor substitution is technologically accompanied by higher capital intensity. Based on these estimates, we verify the sufficient condition for an interior solution derived in Appendix A: for all province-industry observations in the sample, the minimum ratio of the left-hand side to the right-hand side of the condition is 2.7, with a median of 8.3, well above unity. This confirms that the continuous and monotonic $\rho^*(\omega)$ relationship predicted by Lemma 1 is strictly satisfied in the data.

The essence of this identification strategy is to use the persistent cross-province differences in factor prices-induced by China's *hukou* and capital-market segmentation-to trace out the shape of the technology menu. This logic parallels Caselli and Coleman (2006), who used cross-country factor endowment

differences to identify the world technology frontier, but is applied here for the first time to within-country spatial variation.

C. Calibration of the Spatial Friction Parameters

1. Calibration

The theoretical model treats the *hukou* migration cost μ_{ij} and the regional capital wedge τ_i^K as exogenously given institutional features (Section 3). We endow them with operational values in the following manner.

***Hukou* migration frictions** μ_{ij} are defined as the proportional utility discount borne by a worker registered in i and working in j. We combine the cost estimates of Tombe and Zhu (2019) derived from interprovincial migration flows with the empirical evidence in Poncet (2006) and Chan (2010) that *hukou* barriers increase with administrative border and geographic distance. The μ_{ij} are calibrated as a function of distance and a cross-province indicator. In the baseline calibration, the mean μ for intra-provincial migration across prefectures is 1.12, the median for interprovincial migration is about 1.38, and the cost for migration from the western interior to eastern coastal megacities exceeds 1.60. These values are consistent with the documented magnitudes of the public-service discounts faced by rural migrant workers.

Regional capital wedges τ_i^K. In the model, the rental price of capital in a region satisfies $r_i = \tau_i^K \bar{r}$, where $\bar{r}$ is the national benchmark interest rate (taken as the cross-province average of r_i in each year). We therefore directly set $\hat{\tau}_i^K = r_i/\bar{r}$. The estimates reveal that $\hat{\tau}_i^K$ generally lies between 1.05 and 1.20 for coastal provinces, while it ranges between 1.30 and 1.40 for interior provinces, with the highest values concentrated in central and western regions. This pattern is highly consistent with the empirical findings of Boyreau-Debray and Wei (2005) and Song, Storesletten, and Zilibotti (2011) that China's state-dominated banking

system allocates credit preferentially to coastal and state-owned sectors. In general equilibrium, this wedge corresponds to the real resource cost absorbed by financial intermediation; its accounting treatment follows exactly the setup of Section 3, with the wedge revenue distributed as equal per-capita dividends to all workers.

2. Calibration and Estimation Summary

For convenience, we collect in Table 5 all the parameters that are externally calibrated or structurally estimated. Each entry reports the value, the target or source, and the corresponding model equation or proposition.

Table 5: Calibration and Estimation Summary

Notes: Panel A parameters are structurally estimated from the first-order condition (8) and the factor share equation (9). Standard errors in parentheses are clustered at the province level. Ranges in Panels B-D reflect variation across provinces, industries, or robustness specifications. All parameters are held fixed across the baseline and counterfactual simulations, except where explicitly altered (e.g., $\mu_{ij} = 1$, $\tau_i^K = 1$ in Scenarios C).

The parameters listed in Table 5 are exactly those required to solve the general equilibrium and to implement the variance decomposition of firm-level TFPR. The **technology menu parameters** θ_0 and θ_1 capture the theoretically central trade-off that technologies with a higher substitution elasticity are more capital-intensive, and their estimates strictly satisfy the interior solution condition of Lemma 1 for every Chinese province in our sample. The **spatial friction parameters** μ_{ij} and τ_i^K embody the *hukou* system and capital-market segmentation that prevent factor price equalization, thereby generating the persistent cross-region variation in ω_i that triggers endogenous technology choice. The **firm-level wedges** capture residual capital and labor misallocation within

provinces, and their normalization ensures that only the cross-firm dispersion is reflected in the decomposition, not the average level.

This compact summary table, together with the estimation details in Sections 9.1-9.3 and the appendices, provides full replicability and a clean mapping from the theoretical model to the quantitative analysis.

D. Construction of Firm-Level Distortions and TFPR

1. Firm Production Function and Measurement of Distortions

With the regional optimal technology $\hat{\rho}_i^*$ and the corresponding capital distribution parameter $\hat{\gamma}_i = \hat{\gamma}(\hat{\rho}_i^*)$ calibrated, we write the actual production technology of firm f (located in province i, sector s) strictly following the production function formulation of Section 5.1:

$$Y_f = Z_f[\hat{\gamma}_i K_f^{\hat{\rho}_i^*} + (1-\hat{\gamma}_i)L_f^{\hat{\rho}_i^*}]^{1/\hat{\rho}_i^*},$$

where Z_f is the firm-specific Hicks-neutral productivity. Firms face two types of heterogeneous distortions: a capital distortion τ_f^K and a labor distortion τ_f^L, which respectively scale the effective prices of capital and labor to $(1+\tau_f^K)r_i$ and $(1+\tau_f^L)w_i$. Using the first-order condition for factor demand derived in Section 5.2, the optimal capital-labor ratio satisfies:

$$\frac{K_f}{L_f} = \frac{\hat{\gamma}_i}{1-\hat{\gamma}_i}\left[\frac{w_i(1+\tau_f^L)}{r_i(1+\tau_f^K)}\right]^{\frac{1}{1-\hat{\rho}_i^*}}.$$

From equation (12), the observed K_f/L_f, the calibrated regional factor prices, and the technology parameters are sufficient to identify the relative distortion $\ln(1+\tau_f^K) - \ln(1+\tau_f^L)$. To pin down the absolute levels, we adopt HK's approach of normalizing the means of $\ln(1+\tau_f^K)$ and $\ln(1+\tau_f^L)$ to zero within each province-industry cell-equivalent to assuming that, on average, firm-level distortions cancel out within the region and do not affect the mean TFPR. This

normalization has no effect on the measurement of productivity dispersion; it only standardizes the estimated levels.

2. Derivation of the Structural Form of TFPR

Under the assumption of constant markups from monopolistic competition (Section 6), the firm's product price is $P_f = \mu_s \cdot c_f$, where c_f is the unit cost and $\mu_s \equiv \sigma_s/(\sigma_s - 1)$ is the industry-specific constant markup. Revenue total factor productivity (TFPR) is defined as:

$$\text{TFPR}_f \equiv \frac{P_f Y_f}{K_f^{\tilde{\gamma}} L_f^{1-\tilde{\gamma}}},$$

where $\tilde{\gamma}$ is the benchmark capital elasticity used for comparability. To make the results directly comparable to HK's classic accounting, we set $\tilde{\gamma}$ to the national average capital output elasticity $\bar{\gamma} = 0.45$ in the baseline. After algebraic manipulation and substituting the cost function and factor distortions, the log TFPR admits the structural decomposition (see Section 5.2 and supplementary derivations):

$$\ln\text{TFPR}_f = \underbrace{\ln\mu_s}_{\text{industry constant}} + \underbrace{\ln\Xi(\omega_i; \hat{\rho}_i^*)}_{\text{endogenous technology choice component}} + \ln Z_f + \underbrace{F(\tau_f^K, \tau_f^L; \hat{\rho}_i^*)}_{\text{firm distortion component}}.$$

The function $\Xi(\omega_i; \hat{\rho}_i^*)$ is entirely determined by the regional factor prices and the optimal technology parameters; it is a deterministic term common to all firms located in the same province. Its presence means that even if we assume that all firms face zero individual distortions ($\tau_f^K = \tau_f^L = 0$), the TFPR of Chinese manufacturing firms would still display systematic dispersion across provinces because of the differences in ω_i and the corresponding ρ_i^* . According to Proposition 3, such dispersion is a feature of the constrained-Pareto-optimal allocation given existing *hukou* and capital frictions, and theoretically should not be classified as misallocation.

E. Nested Decomposition of Firm-Level TFPR Variance

1. Decomposition Method and Procedure

Following the variance decomposition logic of Proposition 2, we decompose the total variance of firm log TFPR into between (province-industry) and within (firm within province-industry) components:

$$Var(\ln\text{TFPR}_f) = Var(\ln\overline{\text{TFPR}}_{is}) + \mathbb{E}[Var(\ln\text{TFPR}_f|i,s)].$$

We further split the between-group variance into two main sources using the structural function $\ln\Xi$:

Endogenous technology choice contribution: measured directly by $Var[\ln\Xi(\omega_i;\hat{\rho}_i^*)]$, which captures the systematic productivity differences arising from factor endowment variation and the constrained-optimal choice of technology.

Residual regional misallocation: the remainder of the between-group variance that cannot be explained by $\ln\Xi$, computed as $Var(\ln\overline{\text{TFPR}}_{is}) - Var[\ln\Xi]$. This corresponds to structural distortions operating at the regional level that are not directly captured by the measured *hukou* and capital wedges (e.g., province-specific administrative interventions, local protectionism).

The within-group variance is a mixture of **firm-specific distortions** and **exogenous productivity**. To separate them further, we regress firm lnTFPR on $\ln\Xi$, industry fixed effects, and a set of firm characteristics (log assets and its square, ownership type dummies, firm age). The residual from this regression is treated as the exogenous productivity Z_f plus noise. The portion of the within-group variance explained by the characteristic variables, together with the contribution of the distortion function F, is attributed to **firm-specific distortions**; the variance of the residual is designated as **exogenous productivity and**

residual. This conservative procedure ensures that the technology choice component is not contaminated by firm characteristics.

2. Decomposition Results

Table 3 reports the variance decomposition based on the national uniform capital elasticity $\tilde{\gamma} = \bar{\gamma}$.

Table 3: Structural Decomposition of Firm-Level ln **TFPR Variance**

The data show that more than one-quarter of TFPR dispersion originates from cross-regional factor price differences and the associated optimal technology adjustment. In the HK accounting framework, this portion would be implicitly assumed to stem from removable factor price wedges and would be counted toward the potential TFP gains. Our structural decomposition, however, reveals that it is precisely the rational production response to the existing spatial frictions, and has no simple policy handle for efficiency improvement.

If we replace the benchmark capital elasticity $\tilde{\gamma}$ with each province's own $\hat{\gamma}_i$, the variance contribution of endogenous technology choice rises slightly to 28.1%, with within-firm distortions correspondingly edging down. The overall pattern is stable. This directly corroborates the corollary of Proposition 2: whenever one ignores the endogenous regional differences in the production function, a substantial amount of normal productivity dispersion is misdiagnosed as misallocation.

F. Direct Comparison with HK: Multi-Scenario Counterfactual Simulations

To further quantify the welfare contributions of the different types of frictions and to contrast with HK's single accounting, we design three progressively deeper counterfactual scenarios based on the estimated structural parameters, and solve for the new equilibria using the general equilibrium model.

Scenario A (HK-style accounting). Assume that firms in all regions share an identical production function (i.e., $\rho_i^* = \rho$ everywhere, ignoring technology choice differences), and simultaneously eliminate all firm-level capital and labor distortions ($\tau_f^K = \tau_f^L = 0$), reallocating capital and labor within industries until marginal products are equalized. This is equivalent to compressing all observable TFPR dispersion (aside from the exogenous productivity component) to zero. This scenario reproduces the potential gains under HK's rule; we estimate the manufacturing aggregate TFP gain to be **+43.2%**, well within the 30%-50% range reported by HK.

Scenario B (Eliminate firm distortions only, retain spatial technology differentiation). Maintain the calibrated *hukou* migration costs μ_{ij} and regional capital wedges τ_i^K, and only set $\tau_f^K = \tau_f^L = 0$ (or equal to a common intercept) for all firms within each province. Capital and labor are allowed to move freely within provinces but cross-province mobility remains restricted by the original frictions. Firms continue to choose the optimal ρ_i^* according to the local equilibrium ω_i. By Proposition 3, this state is exactly the constrained-optimal allocation given the institutional frictions; the remaining cross-province TFPR differences contain an irreducible endogenous technology-choice component. In this scenario, the TFP gain drops sharply to **+24.9%**.

Scenario C (Complete removal of spatial frictions and firm distortions). Set $\mu_{ij} = 1$, $\tau_i^K = 1$, and simultaneously eliminate firm-level wedges. Labor and capital can move freely across the entire country, leading to factor price equalization. All regions eventually adopt a single optimal technology ρ, achieving the frictionless first-best benchmark described in Section 2 of the theory. The manufacturing TFP gain now reaches **+51.6%**.

Table 4 summarizes the output effects of each scenario.

Table 4: Counterfactual Simulations of Potential Gains in Chinese Manufacturing TFP

The 18.3 percentage point gap between Scenario B and Scenario A cleanly measures the overestimation of the allocative gains in the HK accounting caused by ignoring endogenous technology choice. Strikingly, once the deep spatial institutional barriers are fundamentally lifted, the potential gain is not compressed but actually exceeds the original HK number, because cross-regional factor reallocation induces technological convergence and unlocks additional dynamic efficiency gains omitted in traditional accounting.

G. Robustness Analysis

To verify the reliability of the core conclusions, we conduct a series of robustness checks, each involving a complete re-estimation of the technology menu, the variance decomposition, and the counterfactual simulations.

Functional form of the technology menu. Re-estimating with a logistic form $\gamma(\rho) = e^{\lambda\rho}/(1 + e^{\lambda\rho})$ yields an endogenous technology choice variance share of 25.4% and a Scenario B TFP gain of +26.1%, highly consistent with the baseline.

Markup treatment. Using province-industry average markups instead of firm-specific markups gives a technology choice variance share of 27.2%, with unchanged conclusions.

Subsample analysis. Splitting the sample into coastal and inland provinces reveals that inland provinces, where factor price dispersion is larger, have a technology choice variance share of 30.5%; for coastal provinces it is 21.3%, still a significant portion.

Excluding the WTO accession shock period. Removing the 2001-2003 observations raises the technology choice share to 29.0%, indicating that trade liberalization temporarily compressed interprovincial factor price gaps, and our baseline is likely conservative.

Sensitivity to the migration elasticity. Varying the migration elasticity κ in the theoretical model from 1.5 to 3.5 moves the endogenous technology choice variance share between 23% and 30%, with the Scenario B gain always in the +22%-+27% range.

These tests uniformly demonstrate that the paper's central finding-that more than a quarter of firm-level TFPR dispersion is due to endogenous technology choice adapting to local factor endowments-is robust to reasonable parameter and data perturbations.

H. Economic Implications and Dialogue with the Literature

The quantitative results generate clear academic value added on three levels.

First, they **revise the accounting basis of the misallocation literature** (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). We identify and quantify that when persistent spatial factor price differences exist, observed TFPR dispersion is not entirely equivalent to removable allocative inefficiency. Removing the endogenous technology choice component significantly reduces the potential output gains achievable purely by correcting firm-level factor price wedges. This provides a constrained-efficiency, micro-theoretic correction to misallocation accounting.

Second, they **extend the empirical frontier of the appropriate technology hypothesis** (Basu and Weil, 1998; Caselli and Coleman, 2006). This paper is the first to transplant the technology choice analysis from cross-country comparisons to within-country spatial variation, using China's unique institutionally induced spatial segmentation to detect and measure the contribution of environment-appropriate technology to productivity dispersion at the firm micro level. Our evidence directly supports the theoretical prediction of Atkinson and Stiglitz (1969) that different factor price environments induce different production technologies.

Third, they **deepen the understanding of the aggregate effects of spatial frictions** (Tombe and Zhu, 2019; Caliendo, Dvorkin, and Parro, 2019). Traditional spatial-trade models usually fix the production function and discuss only factor misallocation costs. This paper reveals and quantifies an additional channel: spatial frictions, by forcing each region to adopt a locally idiosyncratic technology system, cause fragmentation of the production structure and thereby impose an extra aggregate productivity loss. When all spatial segmentation is removed, the extra gain from technology convergence (the roughly 8.4 percentage points by which Scenario C exceeds Scenario A) represents the quantitative expression of this channel.

The policy implication is sharp: focusing reform efforts on eliminating inter-firm factor price gaps without touching the fundamental spatial barriers of the *hukou* system and the financial system will capture only a small fraction of the potential efficiency improvement. The correct path to substantially releasing manufacturing productivity is to carry out deep institutional reforms that allow labor and capital to flow freely across regions, thereby simultaneously achieving optimal factor allocation and technological convergence and upgrading.

V. Fiscal Policy in a Second-Best Environment: Theory and Quantitative Evidence

The preceding analysis established that the decentralized equilibrium under spatial frictions is constrained-efficient: given the hukou migration costs μ_{ij} and the capital-market wedges τ_i^K, no policy intervention that respects these frictions can improve upon the market allocation. This result, grounded in the classic theory of the second best (Lipsey and Lancaster, 1956), carries a direct but underexplored implication for fiscal policy. In an economy where factor prices are prevented from equalizing by deep institutional barriers, standard tax and subsidy

instruments that would be purely redistributive or neutral in a first-best world acquire allocative consequences through the endogenous technology choice margin. A capital subsidy targeted at a labor-abundant province, a payroll tax rebate designed to boost manufacturing employment, or a place-based output subsidy intended to attract investment-all alter the effective factor price ratios that firms face, and thereby shift their chosen production functions along the technology menu characterized by $\gamma'(\rho) > 0$. This section provides a formal analysis of this mechanism, derives its welfare implications, and quantifies its empirical significance using the calibrated model.

To introduce fiscal policy into the framework, we extend the model of Sections 2-9 by allowing the government to levy a set of region-specific ad valorem taxes and subsidies on factor inputs and output. Let τ_i^L denote a proportional subsidy to labor employed in region i, so that the after-subsidy wage faced by firms is $(1 - \tau_i^L)w_i$. Let $\tau_i^{K,g}$ denote a proportional tax on capital used in region i (distinct from the institutional capital wedge τ_i^K, which we now relabel $\tau_i^{K,\text{inst}}$ for clarity), so that the after-tax rental rate is $(1 + \tau_i^{K,g})\tau_i^{K,\text{inst}}\overline{r}$. The government finances these interventions through a uniform lump-sum tax (or transfer) T on all workers, ensuring budget balance in the aggregate. The effective wage-rental ratio confronting firms in region i becomes

$$\tilde{\omega}_i \equiv \frac{(1 - \tau_i^L)w_i}{(1 + \tau_i^{K,g})\tau_i^{K,\text{inst}}\overline{r}}.$$

Crucially, the firms technology choice is made after observing this policy-distorted factor price ratio. The firm selects $\tilde{\rho}_i$ to minimize the unit cost function $\mathcal{C}(\tilde{\omega}_i; \rho)$, yielding the first-order condition $\partial \ln \mathcal{C}(\tilde{\omega}_i; \rho)/\partial \rho = 0$. Let $\tilde{\rho}_i = \rho^*(\tilde{\omega}_i)$ denote the optimal technology under the policy regime, and let $\rho_i^* = \rho^*(\omega_i)$ denote the technology that would be chosen in the laissez-faire constrained optimum,

where $\omega_i = w_i/(\tau_i^{K,\text{inst}}\bar{r})$ is the true social factor price ratio. Whenever $\tau_i^L \neq \tau_i^{K,g}$, the effective ratio $\tilde{\omega}_i$ differs from ω_i, and consequently $\tilde{\rho}_i \neq \rho_i^*$.

The welfare consequences of this technology distortion can be characterized analytically. Consider a constrained planner who must respect the deep institutional frictions μ_{ij} and $\tau_i^{K,\text{inst}}$, but who can choose whether to deploy the fiscal instruments τ_i^L and $\tau_i^{K,g}$. The planners problem is to maximize the utilitarian social welfare function defined in Section 7.2, subject to the same resource, technology, and institutional constraints as before, but now with the additional policy variables and the government budget constraint. Proposition 4 establishes that the planner would optimally set all distortionary fiscal instruments to zero.

Proposition 4 (Optimality of Laissez-Faire under Institutional Frictions). Let $\{\tau_i^L, \tau_i^{K,g}\}$ be any set of region-specific taxes and subsidies satisfying aggregate budget balance. The constrained-efficient allocation that solves the planners problem with these instruments available coincides with the laissez-faire constrained optimum; that is, the optimal policy is $\tau_i^L = \tau_i^{K,g} = 0$ for all i.

Proof. The planner maximizes welfare by choosing allocations and policy instruments subject to the constraints. The first-order condition with respect to the technology choice ρ_i is identical to that of the decentralized firm only when the policy-distorted factor price ratio $\tilde{\omega}_i$ coincides with the true social ratio ω_i. If $\tilde{\omega}_i \neq \omega_i$, the firms chosen $\tilde{\rho}_i$ does not minimize the true social unit cost $\mathcal{C}(\omega_i; \rho)$. The planner internalizes this distortion: by setting $\tau_i^L = \tau_i^{K,g} = 0$, the planner eliminates the wedge between private and social factor costs, restoring the alignment between the firms technology choice and the minimizer of true resource costs. Since the planner can achieve any desired redistribution through the lump-sum instrument T without distorting factor prices, the optimal use of

distortionary instruments is zero. A formal derivation, which parallels the proof of Proposition 3 in Appendix E, is provided in a supplementary note. ■

Proposition 4 is the fiscal-policy analogue of the constrained-efficiency result established in Proposition 3. It implies that any regionally differentiated tax or subsidy-whether a capital subsidy for inland provinces, a payroll tax rebate for manufacturing, or an output subsidy for high-tech sectors-reduces aggregate welfare relative to the laissez-faire constrained optimum given the existing spatial frictions. The mechanism is distinct from the standard misallocation channel emphasized by HK. In the HK framework, a capital subsidy that reduces the dispersion of marginal revenue products across firms within an industry would raise aggregate TFP by improving allocative efficiency. In our framework, a capital subsidy targeted at a labor-abundant province alters the effective wage-rental ratio, inducing firms to adopt a more capital-intensive technology than is optimal given the provinces true factor scarcity. The resulting increase in unit costs, evaluated at social factor prices, reduces aggregate output even if it appears to reduce measured TFPR dispersion. The policy thus induces a "technology distortion"-a shift of the production function itself away from the constrained-optimal point on the technology menu.

The economic intuition for this result rests on the fundamental complementarity embedded in the technology menu, $\gamma'(\rho) > 0$. When the government subsidizes capital in a labor-abundant region, it lowers the effective rental rate faced by firms, raising $\tilde{\omega}_i$. Firms respond by selecting a higher $\tilde{\rho}_i$, which, through the menu, implies a more capital-intensive production function. However, this higher ρ_i is socially suboptimal because the true social cost of capital in that region-reflected in the institutional wedge $\tau_i^{K,\text{inst}}$ -remains high. The subsidy induces firms to behave as if capital were cheaper than it truly is, leading them to adopt a production process that economizes on labor-the abundant factor-rather than

capital-the scarce factor. This mismatch between the chosen technology and the regions true factor endowments raises the unit cost of production, and the resulting output loss cannot be compensated by the redistributive gains of the subsidy because the planner could have achieved those same redistributive goals through non-distortionary lump-sum transfers.

To quantify the magnitude of this technology distortion, we conduct a policy simulation using the calibrated model. We consider a stylized place-based policy: a 5 percent wage subsidy ($\tau_i^L = 0.05$) applied to all inland provinces (those west of the Hu Huanyong line, which approximately separates Chinas densely populated coastal regions from its sparsely populated interior). The subsidy is financed by a uniform lump-sum tax on all workers, ensuring aggregate budget balance. We solve for the new general equilibrium under this policy regime, allowing factor prices, migration flows, trade shares, and technology choices to adjust. The equilibrium is computed using the nested fixed-point algorithm described in Appendix I, with the policy-distorted effective wage-rental ratio replacing the laissez-faire ratio in the firms technology choice problem.

Table 6 reports the results. Aggregate manufacturing TFP declines by 1.8 percent relative to the laissez-faire constrained optimum. This output loss is not attributable to the standard Harberger triangle-which, given the 5 percent subsidy rate and empirically plausible labor supply elasticities, would generate a deadweight loss well under 0.5 percent of aggregate output-but is dominated by the technology choice distortion. The subsidy shifts the average ρ_i^* in the targeted inland provinces downward by approximately 0.07 standard deviations relative to the laissez-faire optimum. Because these provinces are already labor-abundant and capital-scarce, their laissez-faire technology choices already feature relatively low substitution elasticities and low capital intensities. The labor subsidy further reduces the effective cost of labor, lowering $\tilde{\omega}_i$ and pushing firms toward even

lower ρ_i^* and less capital-intensive production functions. The resulting unit cost increases, evaluated at the true social factor prices, account for roughly three-quarters of the aggregate TFP loss. The cross-sectional variance of $\ln T(\omega_i)$ widens by approximately 5 percent, as the subsidy introduces additional heterogeneity in the effective factor price ratios faced by firms in different provinces.

The distributional consequences are also revealing. Employment in the subsidized inland provinces rises by approximately 2.1 percent, as the lower effective labor cost encourages labor-intensive production. However, measured TFPR in these provinces falls by 1.4 percent, because the technology shift toward lower ρ_i^* raises unit costs. In the unsubsidized coastal provinces, employment contracts slightly as labor is drawn inland, and TFPR rises modestly. The aggregate effect is a net reduction in real output, concentrated in the very regions the policy was designed to assist. This pattern illustrates a general principle: in a second-best environment where the production function itself responds to policy-altered factor prices, place-based subsidies can be self-defeating. The technology choice margin amplifies the efficiency cost of the intervention while attenuating its intended redistributive benefits.

We also simulate a capital subsidy of equivalent magnitude ($\tau_i^{K,g} = -0.05$) targeted at the same inland provinces. The qualitative effects are symmetric: the subsidy lowers the effective rental rate, raises $\tilde{\omega}_i$, and shifts technology choice toward higher ρ_i^* and more capital-intensive production functions. Aggregate TFP falls by 2.1 percent, a slightly larger loss than under the labor subsidy because the capital subsidy directly counteracts the signal sent by the institutional capital wedge $\tau_i^{K,\text{inst}}$, which already makes capital expensive in inland regions. The subsidy induces firms to adopt technologies that are even more capital-intensive

than those that would be socially optimal given the true cost of capital, compounding the misallocation.

These quantitative findings carry direct implications for the design of regional policy in China and other developing economies with significant internal factor-market frictions. The standard prescription of the misallocation literature-equalize marginal revenue products across firms and regions-implicitly assumes that the production function is invariant to the policy instruments used to achieve that equalization. Our analysis demonstrates that this assumption is not innocuous when technology is endogenous. A capital subsidy that reduces the cross-regional dispersion of measured marginal revenue products may simultaneously shift the underlying production functions in a socially costly direction. The appropriate metric for policy evaluation is not the observed dispersion of TFPR-which confounds efficient technological differentiation with genuine misallocation-but rather the level of aggregate welfare, which accounts for both the direct allocative effects of the policy and the indirect technology choice effects.

Our results also shed light on a long-standing debate regarding the efficacy of place-based policies. Glaeser and Gottlieb (2008) argue that spatially targeted subsidies often fail to raise aggregate welfare because they displace economic activity without increasing overall productivity. Kline and Moretti (2014) provide evidence that certain place-based interventions can raise local productivity, but their analysis takes the production function as given. Our framework identifies an additional mechanism through which place-based subsidies can prove counterproductive: by altering local factor prices, they change the locally optimal production technology, and the direction of this change need not align with the regions comparative advantage. A wage subsidy in a labor-abundant region may seem natural from the perspective of standard trade theory, but if the subsidy distorts technology choice toward an excessively labor-intensive production

function-beyond what is optimal given the regions institutional constraints on capital mobility-the net welfare effect may be negative.

The fiscal policy analysis also contributes to the broader public finance literature on taxation in the presence of pre-existing distortions. It has long been recognized that the welfare cost of a new tax depends on the existing structure of distortions in the economy (Ballard, Shoven, and Whalley, 1985). In the context of our model, the existing distortions are the hukou migration costs and the capital-market wedges, which generate the spatial factor price dispersion that triggers endogenous technology differentiation. Introducing a new fiscal distortion on top of these pre-existing ones creates a double inefficiency: it misallocates factors for a given technology, and it shifts the technology itself away from the constrained-optimal point. The interaction between the two layers of distortion implies that the marginal excess burden of a new tax is higher in our framework than in a model with an exogenous production function. This amplification channel is not captured by standard computable general equilibrium models that hold production parameters fixed across regions. Our quantitative results suggest that the technology choice margin can multiply the welfare cost of a given fiscal intervention by a factor of three to four relative to the standard Harberger calculation.

Finally, the analysis speaks to the literature on fiscal federalism and intergovernmental transfers in China. A large empirical literature documents that Chinas system of fiscal transfers from the central government to provinces has historically favored interior and western regions (Tsui, 2005; Zhang, 2006). To the extent that these transfers operate as implicit subsidies that lower the effective cost of capital or labor in recipient provinces, they alter the local wage-rental ratios that drive technology choice. Our framework suggests that the welfare consequences of such transfers depend not only on their direct allocative effects-which the standard misallocation literature would capture-but also on their

indirect effects through the endogenous technology margin. A transfer that effectively subsidizes capital in a labor-abundant province would push that province toward a more capital-intensive technology, potentially raising measured TFP but moving the province away from the production function that minimizes true social costs given its factor endowments. The normative evaluation of Chinas transfer system, which has traditionally focused on the equity-efficiency trade-off, may therefore need to incorporate this additional channel.

The central message of this section is that in an economy characterized by persistent spatial frictions that prevent factor price equalization, the policy-maker operates in a genuinely second-best environment. Fiscal instruments that would be neutral or unambiguously welfare-improving in a first-best world become potential sources of distortion through their effects on the technology choice margin. The appropriate policy response is not to multiply taxes and subsidies in an effort to offset the symptoms of factor-market segmentation-measured TFP dispersion, regional wage gaps, or variations in capital intensity-but rather to target the underlying institutional barriers themselves. Hukou reform and financial liberalization, by reducing spatial factor price dispersion at its source, would eliminate both the direct allocative inefficiencies emphasized by HK and the technology choice distortions we identify, allowing Chinas manufacturing sector to converge not only in factor allocation but in the very production functions it employs.

VI. Concluding Remarks

This paper began with a puzzle that has long occupied economists studying Chinas remarkable growth experience: how can an economy sustain rapid aggregate productivity gains while simultaneously exhibiting enormous, persistent dispersion in measured total factor productivity across its regions? The standard

misallocation framework, pioneered by Hsieh and Klenow (2009), interprets this dispersion as evidence of distortionary wedges that prevent the efficient allocation of productive factors, and calculates that eliminating these wedges would raise Chinese manufacturing TFP by 30 to 50 percent. This diagnosis rests, implicitly but fundamentally, on the assumption that a single production function is equally appropriate for all firms in all locations. Once that assumption is relaxed, the interpretation of observed TFP dispersion changes in ways that matter both for measurement and for policy.

We have developed a spatial general equilibrium model in which the production function itself is an equilibrium object, optimally selected by firms from a continuous menu of techniques in response to the local factor prices they face. When institutional frictions-the hukou household registration system and capital-market segmentation-prevent factor price equalization across regions, firms in different provinces optimally choose different production functions. Those facing high wage-rental ratios adopt more flexible, capital-intensive technologies; those facing abundant labor and expensive capital choose more rigid, labor-intensive ones. This mechanism generates what we have termed *endogenous TFP heterogeneity*: cross-regional differences in measured productivity that arise not from misallocation but from firms optimal adaptation to the constraints imposed by the institutional environment.

The theoretical framework yields three central results. First, in a frictionless benchmark where labor and capital are perfectly mobile, factor prices equalize, all regions adopt identical technologies, and measured TFP dispersion reduces to the variance of exogenous productivity differences alone-confirming that standard misallocation accounting implicitly assumes a frictionless world. Second, in the presence of spatial frictions, the cross-sectional variance of measured TFP admits an exact decomposition into the variance of exogenous productivity, the variance of the endogenous technology choice term, and their covariance. The technology

choice component does not represent a deviation from an attainable efficient allocation; it is a feature of the constrained-efficient, second-best equilibrium, as we demonstrate by solving a constrained planners problem that respects the same spatial frictions as the market. Third, fiscal policy interventions that alter relative factor prices-regional capital subsidies, payroll tax rebates, or place-based output incentives-become sources of distortion through the technology choice margin. Because the production function itself responds to policy-altered factor prices, instruments that would be neutral in a first-best world can reduce aggregate welfare in the second-best environment created by spatial frictions.

The quantitative analysis, calibrated to the same Chinese manufacturing data that underpin the original HK estimates, reveals that the mechanism is not merely a theoretical possibility but a first-order empirical phenomenon. Endogenous technology choice accounts for approximately 27 percent of the total firm-level TFPR variance in Chinese manufacturing. A counterfactual simulation that replicates the HK calculation-assuming a common production function across all firms and removing firm-level distortions-yields a potential TFP gain of 43 percent, consistent with the original estimates. However, when we correct this calculation to account for the fact that firms in different regions have optimally chosen different technologies, the gain from removing firm-specific distortions falls sharply to 25 percent. The 18 percentage point difference is the overestimation introduced by treating all TFPR dispersion as misallocation. An alternative counterfactual that removes all spatial frictions-the hukou system and capital-market segmentation, alongside firm-level distortions-yields a gain of 52 percent, exceeding the original HK estimate because factor mobility induces not only optimal reallocation but also technological convergence across regions.

Our analysis also demonstrates the potentially counterproductive effects of place-based fiscal policies in this second-best environment. A 5 percent wage subsidy targeted at inland provinces, financed by a uniform lump-sum tax,

reduces aggregate manufacturing TFP by 1.8 percent relative to the laissez-faire constrained optimum, with technology choice distortions accounting for roughly three-quarters of this loss. A capital subsidy of equivalent magnitude produces a similar welfare reduction. These results carry a clear policy message: in an economy where deep institutional barriers prevent factor price equalization, the welfare-improving reform is to reduce the barriers themselves, not to deploy taxes and subsidies that attempt to manage the symptoms of factor-market segmentation-measured TFP dispersion, regional wage gaps, or variations in capital intensity-while leaving their causes intact.

The paper makes several contributions to the existing literature. To the misallocation literature, we provide a structural framework for distinguishing between genuine allocative inefficiency and the productivity consequences of optimal technological differentiation-a distinction that significantly affects quantitative estimates of the gains from reform. To the appropriate technology literature, we offer the first within-country spatial application, demonstrating that the mechanisms Basu and Weil (1998) and Caselli and Coleman (2006) identified across countries operate with equal force across regions within a single large economy when institutional frictions segment factor markets. To the spatial economics literature, we introduce endogenous technology choice as a new channel through which spatial frictions affect aggregate productivity, complementing the traditional focus on factor misallocation costs.

Methodologically, the paper offers a template for how to take the endogeneity of the production function seriously in quantitative spatial analysis. The decomposition of TFP variance into its exogenous, endogenous technology, and genuine misallocation components is portable to other settings-other countries with significant internal frictions, other time periods, other levels of aggregation-and provides a diagnostic tool for distinguishing between efficient adaptation and correctable inefficiency. The integration of the De Loecker and Warzynski (2012)

markup correction into the structural estimation of the technology menu demonstrates how product market imperfections and factor market frictions can be jointly analyzed within a unified framework.

Several extensions of this work merit further investigation. The analysis has focused on a static menu of technologies; in a dynamic setting, sustained spatial factor price differentials would likely induce shifts in the menu itself, as the directed technical change literature predicts, potentially amplifying the effects we document. The model has assumed that the technology menu is common across provinces; relaxing this assumption to allow for FDI-induced or policy-induced shifts in the frontier would enrich the analysis of how globalization and industrial policy interact with spatial frictions. The normative analysis has been conducted under the assumption that the deep spatial frictions-the hukou system and capital-market segmentation-are exogenous constraints; endogenizing the political economy of these institutions would provide a deeper understanding of why they persist and how they might be reformed.

The central message of this paper is that not all TFP dispersion is created equal. Some of it reflects genuine misallocation that policy can and should correct. But a substantial portion-we estimate more than a quarter in the Chinese context-is the signature of firms making the best of the constraints they face. Confusing the two leads not only to overestimates of what policy can achieve, but to policies that may be actively counterproductive. The path to higher aggregate productivity in economies characterized by deep spatial frictions runs through institutional reform that removes the frictions themselves-hukou liberalization, financial sector reform, and the creation of integrated national factor markets-rather than through interventions that target the productivity differences these frictions generate. When factor markets are freed, technology follows.

REFERENCES

Acemoglu, D. (1998). "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality." Quarterly Journal of Economics, 113(4), 1055-1089.

Acemoglu, D. (2002). "Directed Technical Change." Review of Economic Studies, 69(4), 781-809.

Ackerberg, D. A., K. Caves, and G. Frazer (2015). "Identification Properties of Recent Production Function Estimators." Econometrica, 83(6), 2411-2451.

Allen, T., and C. Arkolakis (2014). "Trade and the Topography of the Spatial Economy." Quarterly Journal of Economics, 129(3), 1085-1140.

Atkinson, A. B., and J. E. Stiglitz (1969). "A New View of Technological Change." Economic Journal, 79(315), 573-578.

Ballard, C. L., J. B. Shoven, and J. Whalley (1985). "General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States." American Economic Review, 75(1), 128-138.

Basu, S., and D. N. Weil (1998). "Appropriate Technology and Growth." Quarterly Journal of Economics, 113(4), 1025-1054.

Boyreau-Debray, G., and S.-J. Wei (2005). "Pitfalls of a State-Dominated Financial System: The Case of China." NBER Working Paper No. 11214.

Brandt, L., J. Van Biesebroeck, L. Wang, and Y. Zhang (2017). "WTO Accession and Performance of Chinese Manufacturing Firms." American Economic Review, 107(9), 2784-2820.

Broda, C., and D. E. Weinstein (2006). "Globalization and the Gains from Variety." Quarterly Journal of Economics, 121(2), 541-585.

Caliendo, L., and F. Parro (2015). "Estimates of the Trade and Welfare Effects of NAFTA." Review of Economic Studies, 82(1), 1-44.

Caselli, F., and W. J. Coleman (2006). "The World Technology Frontier."

American Economic Review, 96(3), 499-522.

Chan, K. W. (2010). "The Household Registration System and Migrant Labor in China: Notes on a Debate." Population and Development Review, 36(2), 357-364.

De Loecker, J., and F. Warzynski (2012). "Markups and Firm-Level Export Status." American Economic Review, 102(6), 2437-2471.

Dixit, A. K., and J. E. Stiglitz (1977). "Monopolistic Competition and Optimum Product Diversity." American Economic Review, 67(3), 297-308.

Eaton, J., and S. Kortum (2002). "Technology, Geography, and Trade." Econometrica, 70(5), 1741-1779.

Foster, L., J. Haltiwanger, and C. Syverson (2008). "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" American Economic Review, 98(1), 394-425.

Glaeser, E. L., and J. D. Gottlieb (2008). "The Economics of Place-Making Policies." Brookings Papers on Economic Activity, 39(1), 155-253.

Hsieh, C.-T., and P. J. Klenow (2009). "Misallocation and Manufacturing TFP in China and India." Quarterly Journal of Economics, 124(4), 1403-1448.

Kline, P., and E. Moretti (2014). "People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs." Annual Review of Economics, 6(1), 629-662.

Lipsey, R. G., and K. Lancaster (1956). "The General Theory of Second Best." Review of Economic Studies, 24(1), 11-32.

Meng, X. (2012). "Labor Market Outcomes and Reforms in China." Journal of Economic Perspectives, 26(4), 75-102.

Poncet, S. (2006). "Provincial Migration Dynamics in China: Borders, Costs and Economic Motivations." Regional Science and Urban Economics, 36(3), 385-398.

Redding, S. J., and E. Rossi-Hansberg (2017). "Quantitative Spatial Economics." *Annual Review of Economics*, 9, 21-58.

Restuccia, D., and R. Rogerson (2008). "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments." *Review of Economic Dynamics*, 11(4), 707-720.

Song, Z., K. Storesletten, and F. Zilibotti (2011). "Growing Like China." *American Economic Review*, 101(1), 196-233.

Tombe, T., and X. Zhu (2019). "Trade, Migration, and Productivity: A Quantitative Analysis of China." *American Economic Review*, 109(5), 1843-1881.

Tsui, K. (2005). "Local Tax System, Intergovernmental Transfers and Chinas Local Fiscal Disparities." Journal of Comparative Economics, 33(1), 173-196.

Zhang, X. (2006). "Fiscal Decentralization and Political Centralization in China: Implications for Growth and Inequality." Journal of Comparative Economics, 34(4), 713-726.

Appendix A: Proof of Lemma 1 and Characterization of Admissible Technology Menus

This appendix provides a rigorous proof of Lemma 1, fully characterizes the class of functions $\gamma(\rho)$ for which the result holds globally, and offers the economic intuition for why $\gamma'(\rho)$ must be sufficiently large.

Lemma 1 (Comparative Statics of Technology Choice). Assume $\gamma'(\rho) > 0$ for all $\rho \in [\underline{\rho}, \bar{\rho}]$, and that at the unique interior optimum $\rho^* \in (\underline{\rho}, \bar{\rho})$ the second-order condition $\partial^2 \ln \mathcal{C}/\partial \rho^2 > 0$ holds. Then $d\rho^*/d\omega < 0$.

Proof. From the main text, the log unit cost component is $\ln \mathcal{C}(\omega; \rho) = -\frac{1-\rho}{\rho} \ln \Phi(\rho, \omega)$, where the auxiliary function is defined as

$$\Phi(\rho, \omega) \equiv \gamma(\rho)^{\frac{1}{1-\rho}} + (1 - \gamma(\rho))^{\frac{1}{1-\rho}} \omega^{-\frac{\rho}{1-\rho}}.$$

The first-order condition $\mathcal{L}_\rho \equiv \partial \ln \mathcal{C}/\partial \rho = 0$ yields

$$\frac{1}{\rho^2} \ln \Phi(\rho, \omega) - \frac{1-\rho}{\rho} \frac{\Phi_\rho(\rho, \omega)}{\Phi(\rho, \omega)} = 0,$$

which implies the relationship $\Phi_\rho/\Phi = \ln \Phi/[\rho(1-\rho)]$ at the optimum. Applying the Implicit Function Theorem to the identity $\mathcal{L}_\rho(\rho^*(\omega), \omega) \equiv 0$ gives

$$\frac{d\rho^*}{d\omega} = -\frac{\mathcal{L}_{\rho\omega}}{\mathcal{L}_{\rho\rho}},$$

where $\mathcal{L}_{\rho\rho} \equiv \partial^2 \ln \mathcal{C}/\partial \rho^2 > 0$ by the second-order condition, and $\mathcal{L}_{\rho\omega} \equiv \partial^2 \ln \mathcal{C}/\partial \rho \partial \omega$. Hence $sgn(d\rho^*/d\omega) = -sgn(\mathcal{L}_{\rho\omega})$.

Direct differentiation of $\mathcal{L}_\rho$ with respect to ω yields

$$\mathcal{L}_{\rho\omega} = \frac{1}{\rho^2} \frac{\Phi_\omega}{\Phi} - \frac{1-\rho}{\rho} \frac{\Phi_{\rho\omega}\Phi - \Phi_\rho \Phi_\omega}{\Phi^2}.$$

Substituting the first-order condition to eliminate Φ_ρ/Φ, we obtain

$$\mathcal{L}_{\rho\omega} = \frac{1-\rho}{\rho}\frac{\Phi_{\rho\omega}}{\Phi} + \frac{\Phi_{\omega}}{\rho^2\Phi}\left(1 + \frac{1-\rho}{\rho}\ln\Phi\right).$$

We now compute the partial derivatives of Φ. First,

$$\Phi_{\omega} = \frac{\partial}{\partial\omega}\left[(1-\gamma)^{\frac{1}{1-\rho}}\omega^{-\frac{\rho}{1-\rho}}\right] = -(1-\gamma)^{\frac{1}{1-\rho}}\frac{\rho}{1-\rho}\omega^{-\frac{1}{1-\rho}} < 0,$$

for $\rho \in (0,1)$. Second, $\Phi_{\rho\omega} \equiv \partial\Phi_{\omega}/\partial\rho$. A careful computation, provided in the Supplementary Appendix, shows that $\Phi_{\rho\omega} > 0$ whenever $\gamma'(\rho)$ is sufficiently large relative to the curvature of γ. Intuitively, a higher ρ makes the function Φ more sensitive to changes in ω because it raises the exponent on the factor price ratio and tilts the capital share upward. The envelope properties of the minimization problem guarantee that the term $(1 + \frac{1-\rho}{\rho}\ln\Phi)$ remains bounded, so its contribution does not overturn the sign of the first term in $\mathcal{L}_{\rho\omega}$.

A sufficient condition for $\mathcal{L}_{\rho\omega} > 0$ is

$$\frac{\gamma'(\rho)}{\gamma(\rho)} > \frac{1}{(1-\rho)^2}\left[1 + \omega^{\frac{\rho}{1-\rho}}\left(\frac{\gamma}{1-\gamma}\right)^{\frac{1}{1-\rho}}\right]^{-1}.$$

This condition has a clear economic interpretation. The left-hand side is the elasticity of the capital distribution parameter with respect to the substitution parameter: it measures how strongly capital intensity responds to increased factor substitutability. The right-hand side is a threshold that depends on the wage-rental ratio ω and on the current capital intensity γ. When capital is already relatively cheap (high ω), the threshold is lower, making the condition easier to satisfy. When γ is extreme (very close to 0 or 1), the threshold rises, reflecting the fact that a production function that is already extremely labor-intensive or extremely capital-intensive offers little room for further adjustment through changes in ρ.

The economic role of this condition merits emphasis. A higher ρ reduces unit costs through two distinct channels: it raises the elasticity of substitution, which

makes it easier to economize on the factor that has become more expensive; and it increases capital intensity through $\gamma'(\rho) > 0$, which further reduces costs when capital is the relatively cheap factor. The second channel only operates if $\gamma'(\rho)$ is large enough. If $\gamma'(\rho)$ were too small-if more flexible technologies were only marginally more capital-intensive-the firm would gain little from adopting a higher ρ beyond what it could achieve by simply substituting capital for labor along a fixed isoquant. In that case, the firm might optimally choose a corner solution, remaining at $\underline{\rho}$ for all ω below some threshold and jumping to $\bar{\rho}$ above it. The condition above ensures that the interior trade-off is sufficiently steep to generate a continuous, monotone relationship between ρ^* and ω. Importantly, this condition is independent of the market structure: it derives solely from the cost minimization problem and the properties of the technology menu, and therefore applies equally under perfect competition and monopolistic competition.

We now characterize the class of functions $\gamma(\rho)$ for which the sufficient condition holds globally.

Linear specification. $\gamma(\rho) = \theta_0 + \theta_1\rho$, with $\theta_1 > 0$ and θ_0, θ_1 such that $\gamma(\rho) \in (0,1)$ for all $\rho \in [\underline{\rho}, \bar{\rho}]$. For this specification, $\gamma'(\rho)/\gamma(\rho) = \theta_1/(\theta_0 + \theta_1\rho)$, which is decreasing in ρ. The sufficient condition is satisfied for all ρ, ω in the admissible range provided θ_1 is bounded away from zero. With our empirical estimates of $\hat{\theta}_0 = 0.28$ and $\hat{\theta}_1 = 0.34$, the condition is easily met for all Chinese provinces in our sample.

Exponential specification. $\gamma(\rho) = \exp(\lambda\rho)/(1 + \exp(\lambda\rho))$ for $\lambda > 0$. Here $\gamma'(\rho)/\gamma(\rho) = \lambda(1 - \gamma(\rho))$, which is positive and bounded away from zero for interior γ. The sufficient condition holds for all $\lambda > 0$.

Logistic and other S-shaped specifications. Any monotonically increasing function with a bounded derivative that is not too small relative to its level will

satisfy the condition for empirically relevant ranges of ω. Our robustness checks in Section 9 confirm that the main quantitative results are insensitive to the choice among these functional forms.

General class. More generally, let $\gamma(\rho)$ be any continuously differentiable function satisfying $\gamma'(\rho) > 0$ and the curvature condition $\gamma''(\rho) \leq 0$ (concavity). For this class, $\gamma'(\rho)/\gamma(\rho)$ is decreasing, and the sufficient condition holds for all admissible ρ, ω provided $\gamma'(\bar{\rho})/\gamma(\bar{\rho})$ exceeds the maximal value of the right-hand side over the admissible range.

Under any of these conditions, $\mathcal{L}_{\rho\omega} > 0$ and consequently $d\rho^*/d\omega < 0$, which completes the proof. ■

We note that the condition is directly verifiable for each province-year observation in our data. For all province-sector cells in our Chinese estimation sample, the left-hand side of the inequality exceeds the right-hand side by a factor of at least 2.5, confirming that the interior solution is empirically relevant and that the comparative static result applies globally.

Appendix B: Existence of Equilibrium

This appendix provides a constructive proof of existence for the competitive equilibrium defined in Definition 1.

Let $\boldsymbol{\omega} = (\omega_1, ..., \omega_N) \in \mathbb{R}^N_{++}$ be the vector of wage-rental ratios. For any given $\boldsymbol{\omega}$, the following objects are determined uniquely and continuously. First, the optimal technology choice $\rho_i^* = \rho^*(\omega_i)$ is the unique minimizer of $\ln\mathcal{C}(\omega_i; \rho)$ over $[\underline{\rho}, \bar{\rho}]$. Uniqueness follows from the strict convexity of $\ln\mathcal{C}$ in ρ, which is guaranteed by the second-order condition and the characterization of admissible $\gamma(\rho)$ in Appendix A. Continuity of $\rho^*(\omega_i)$ follows from the Theorem of the Maximum. Second, the minimized unit cost component $c_i^* = \mathcal{C}(\omega_i; \rho^*(\omega_i))$ is continuous in ω_i by the same argument. Third, given these unit costs, mill prices are determined by the constant-markup rule $p_{i,s} = \mu_s r_i c_i^* / Z_{i,s}$, where $\mu_s \equiv \sigma_s/(\sigma_s - 1)$. Equilibrium trade shares $\pi_{ij,s}$ are then determined by the gravity equation, which is continuous in the vector of mill prices. Trade shares, in turn, determine regional incomes through the goods market clearing and income identities, all of which are continuous functions of the underlying prices and quantities.

We now construct the excess demand functions. For any $\boldsymbol{\omega}$, the capital rental rate satisfies $r_i = \tau_i^K \bar{r}$, and the wage rate is uniquely determined by $w_i = \omega_i r_i$. Labor demand in each region-sector is derived from the conditional factor demand equation and the output levels required to satisfy final demand, which are smooth functions of prices and incomes. Regional labor demand is obtained by summing across sectors. Labor supply is determined by the migration shares, which are continuous in real incomes and hence in the vector of factor prices.

The excess demand correspondence maps the compact set of possible wage-rental ratio vectors to excess demands for labor and capital in each region. Compactness is achieved by bounding factor prices above zero and below infinity,

which is feasible because endowments and frictions ensure that equilibrium factor prices are bounded away from zero and infinity. All mappings are continuous under the maintained assumptions, including the constant-markup pricing rule. By Kakutanis fixed-point theorem, a fixed point exists at which all factor markets clear simultaneously. This fixed point constitutes a competitive equilibrium.

Global uniqueness is not guaranteed because the system is highly nonlinear due to the endogenous technology choice and the gravity trade structure. However, under standard regularity conditions on the Jacobian of the excess demand system-in particular, gross substitutability in factor demands-the equilibrium is locally unique. A detailed computational algorithm is available in the Supplementary Appendix.

Appendix C: Deterministic Microfoundation for the Technology Menu

This appendix provides a purely deterministic microfoundation for the positive relationship between ρ and γ, replacing the factor price uncertainty mechanism that appeared in earlier versions of this manuscript. The derivation builds on the directed technical change framework of Acemoglu (2002), adapted to a static technology choice setting.

Consider a representative firm that must select a production blueprint before production begins. A blueprint is characterized by a pair of factor-augmenting coefficients (A_K, A_L) and a substitution parameter ρ, with the associated production function

$$Y = Z[\zeta(A_K K)^{\rho} + (1-\zeta)(A_L L)^{\rho}]^{1/\rho},$$

where $\zeta \in (0,1)$ is a fixed parameter. The firm chooses a point on the research possibility frontier

$$A_K^{\psi} + A_L^{\psi} = H(\rho),$$

with $\psi > 1$ and $H'(\rho) < 0$. The parameter ψ governs the curvature of the trade-off between augmenting capital and labor: a larger ψ implies that achieving high efficiency in one dimension requires a greater sacrifice in the other. The function $H(\rho)$ captures the notion that achieving a higher degree of factor substitutability is technically demanding. As ρ increases, the production blueprint must accommodate a wider range of input combinations, which consumes scarce engineering resources. The total factor-augmenting efficiency that can be achieved with a given research budget therefore declines: the frontier shifts inward.

The key difference from the earlier uncertainty-based formulation is that the firm does not face a distribution of future factor prices. Instead, the firm chooses (A_K, A_L, ρ) to minimize the deterministic unit cost given the current local factor prices (w, r). The Lagrangian for this problem is

$$\mathcal{L} = \ln c(A_K, A_L, \rho; w, r) + \lambda(H(\rho) - A_K^{\psi} - A_L^{\psi}),$$

where $\ln c$ is the log unit cost function. The first-order conditions with respect to A_K and A_L yield

$$\frac{A_K^{\psi-1}}{A_L^{\psi-1}} = \frac{\partial \ln c / \partial A_K}{\partial \ln c / \partial A_L}.$$

Evaluating the partial derivatives of $\ln c$, one finds that the right-hand side is increasing in ρ for a given factor price ratio. This reflects a fundamental complementarity: a higher substitution elasticity ρ makes the production process more responsive to factor-augmenting technical improvements in the capital dimension, thereby increasing the marginal benefit of directing research effort toward A_K. The firm optimally biases its blueprint toward capital augmentation.

Solving the first-order conditions yields the optimal augmentation ratio

$$\frac{A_K(\rho)}{A_L(\rho)} = \left(\frac{\zeta}{1-\zeta}\right)^{\frac{1}{\psi-1}} \left[\left(\frac{w}{r}\right)^{-\frac{\rho}{1-\rho}}\right]^{\frac{1}{\psi-1}},$$

which is increasing in ρ whenever $w/r > 1$. Economies where capital is relatively abundant or cheap compared to labor will therefore see a stronger bias toward capital augmentation as substitutability rises.

Substituting the optimal augmenting coefficients into the definition of the capital distribution parameter,

$$\gamma(\rho) = \frac{\zeta A_K(\rho)^{\rho}}{\zeta A_K(\rho)^{\rho} + (1-\zeta) A_L(\rho)^{\rho}},$$

and differentiating with respect to ρ yields $\gamma'(\rho) > 0$ under the maintained assumptions $\psi > 1$ and $H'(\rho) < 0$. The sign is unambiguous: as ρ increases, both the direct substitution effect (higher ρ raises the exponent on capital in the production function) and the induced reallocation of research effort (the firm biases augmentation toward capital) operate in the same direction, raising the effective weight of capital in the production function.

This deterministic microfoundation is fully consistent with the static general equilibrium framework in the main text. Unlike the uncertainty-based formulation in earlier drafts, it does not introduce any stochastic elements that would be absent from the household migration problem or the planners objective. The positive association between ρ and γ emerges as a robust property of optimal resource allocation in research, grounded in the fundamental complementarity between factor substitutability and factor-augmenting technical change. This resolves the internal inconsistency identified by reviewers, in which firms were assumed to optimize against factor price risk that households ignored when making location decisions. The derivation is also entirely independent of the market structure in product markets: it relies only on the firms cost minimization problem given factor prices, which is the same under both perfect and monopolistic competition.

The derivation also clarifies why the firm cares about the substitution elasticity even in a static, deterministic setting. The firm is not indifferent among technologies with different ρ that achieve the same unit cost at the optimal input ratio, because different ρ induce different allocations of research effort across the two augmenting dimensions, and the resulting differences in A_K and A_L affect unit costs even at the optimal factor mix. The substitution elasticity influences the firms research productivity, not merely the shape of the isoquant.

Appendix D: Identification with Monopolistic Competition, Markup Correction, and Menu Heterogeneity

This appendix presents the identification strategy for the technology menu parameters under monopolistic competition, where the observed revenue shares of capital are contaminated by firm-level markups. Unlike earlier versions of this paper, which treated markup correction as a supplementary robustness exercise, we now embed it as an integral and logically necessary component of the identification procedure. This shift reflects the adoption of monopolistic competition as the baseline market structure in the main text.

The fundamental challenge is that under monopolistic competition, the observed revenue share of capital for a firm is $s_{K,i}^R = \gamma_i/\mu_i$, where γ_i is the true output elasticity of capital and $\mu_i > 1$ is the firm-level markup. Using $s_{K,i}^R$ directly in the factor share equation (D2) would therefore conflate variation in the technology parameter $\gamma(\rho)$ with variation in market power. The direction of this bias depends on the spatial correlation between markups and factor prices: if high-wage coastal provinces also exhibit systematically higher markups-a pattern consistent with the evidence in Brandt, Van Biesebroeck, Wang, and Zhang (2017)-then using uncorrected shares would overstate the capital intensity of high-ω regions and bias the estimated $\gamma(\rho)$ curve.

We address this challenge using the control function approach of De Loecker and Warzynski (2012, henceforth DLW). The DLW method exploits the first-order condition for a flexible, undistorted input-materials-to recover firm-level markups without imposing a specific demand structure beyond cost minimization. Under cost minimization, the firm equates the marginal revenue product of materials to the materials price, leading to

$$\mu_i = \frac{\theta_i^M}{s_i^M},$$

where θ_i^M is the output elasticity of materials and s_i^M is the revenue share of materials. The output elasticity θ_i^M is estimated from a production function using the Ackerberg, Caves, and Frazer (2015) correction for simultaneity bias. With consistent estimates of μ_i in hand, the true capital share is recovered as $s_{K,i} = s_{K,i}^R \cdot \mu_i$.

The technology menu parameters are then estimated in two integrated stages. In the first stage, we apply the DLW procedure at the firm level to obtain $\hat{\mu}_i$ and compute the markup-corrected capital shares $\hat{s}_{K,i}$. These corrected shares are aggregated to the province-sector level by taking revenue-weighted averages. In the second stage, we solve the system of the first-order condition $\partial \ln \mathcal{C}(\omega_i; \rho)/\partial\rho = 0$ and the factor share equation

$$s_{K,i} = \frac{\gamma(\rho^*)^{\frac{1}{1-\rho^*}}}{\gamma(\rho^*)^{\frac{1}{1-\rho^*}} + (1-\gamma(\rho^*))^{\frac{1}{1-\rho^*}} \omega_i^{\frac{\rho^*}{1-\rho^*}}}$$

numerically for each province-sector observation, using the corrected shares. This yields a locus of identified pairs $(\hat{\rho}_i^*, \hat{\gamma}_i)$. The linear specification $\gamma(\rho) = \theta_0 + \theta_1\rho$ is then fitted by non-linear least squares.

Using the markup-corrected capital shares, the estimated $\hat{\theta}_1$ is 0.29 (compared to 0.34 that would be obtained from uncorrected shares), and the share of TFPR variance attributable to endogenous technology heterogeneity is 37% (compared to 40% without correction). The modest adjustment reflects an important empirical regularity: markup dispersion, while substantial in the Chinese manufacturing data, is not strongly correlated with the spatial variation in factor prices that drives the technology choice mechanism. Provinces with similar wage-rental ratios exhibit similar average markups, so the cross-sectional variation in capital shares is dominated by the technology choice channel rather than by market power. This finding itself carries economic content: it suggests that

product market power and factor market distortions operate through largely orthogonal channels in Chinas spatial economy.

We also discuss the theoretical implications of imperfect competition for the constrained-efficiency result. Under Dixit-Stiglitz monopolistic competition with a constant elasticity of substitution across varieties, markups are constant and identical across all firms within a sector. In this case, the revenue share $s_{K,i}^{R}$ is proportional to the true output elasticity, and all cross-sectional variation in corrected shares stems from the technology choice mechanism. The constrained-efficiency result of Proposition 3 extends directly to this environment because the planner faces the same CES demand structure and therefore the same constant-markup pricing rule as private firms. The planner cannot eliminate the markup without eliminating the product differentiation that the markup compensates. Under variable markups-arising, for instance, from oligopolistic competition or non-CES demand-the planners allocation may diverge from the market outcome, and the constrained-efficiency claim would need to be qualified. This is an important distinction: our framework isolates the technology choice component of TFP dispersion from the markup dispersion component, and only the latter, when it deviates from the sectoral mean, constitutes genuine misallocation in our framework.

We now state the core identifying assumption and discuss its empirical content. The technology menu $\gamma(\rho)$ is assumed to be common across all regions. This assumption allows us to use cross-regional variation in ω_i and corrected factor shares to trace out a single structural curve. If the menu were region-specific, the locus of identified pairs would confound movement along a menu with shifts of the menu itself. We evaluate this assumption empirically using a regression of corrected capital shares on ω_i with province fixed effects. Under the null of a common menu, the province fixed effects should be jointly insignificant. In our

calibration sample, this null hypothesis is not rejected (p-value = 0.31), providing empirical support for the common-menu assumption.

To further test the robustness of our conclusions, we relax the common-menu assumption by allowing the intercept of the technology menu to depend on observable provincial characteristics. Specifically, we specify

$$\theta_{0,i} = \beta_0 + \beta_1 \cdot FDI_i + \beta_2 \cdot ExportPropensity_i,$$

where FDI_i is the province-level ratio of foreign direct investment to GDP and $ExportPropensity_i$ is the ratio of exports to gross output. The estimated β_1 is positive but small in magnitude (0.07, standard error 0.03), and the share of TFPR variance attributable to endogenous technology heterogeneity declines modestly from 37% to 35%. When we further allow the slope θ_1 to vary with FDI exposure, the interaction term is insignificant. Restricting the sample to inland provinces only yields an endogenous heterogeneity share of 39%, compared to 34% for coastal provinces alone. These results confirm that while coastal-inland differences in the technology frontier are detectable, they do not drive our central finding.

Appendix E: Proof of Proposition 3 (Constrained Efficiency under Monopolistic Competition)

This appendix provides the complete proof that the decentralized equilibrium under monopolistic competition coincides with the solution to the constrained planners problem.

Proposition 3 (Constrained Efficiency). The allocation of the decentralized competitive equilibrium defined in Definition 1 coincides with the solution to the constrained planners problem. That is, the market equilibrium is constrained-efficient given the institutional frictions μ_{ij} and τ_i^K.

Proof. The planner maximizes a utilitarian social welfare function defined as the sum of the expected utilities of all workers. Given the logit choice structure of Section 4.2, the planners objective is

$$W = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{S} m_{i,js}\,\overline{L}_i(\ln y_{i,js} + \ln A_j - \ln \mu_{ij} + \mathbb{E}[\epsilon_{js}|\text{worker chooses } (j,s)]).$$

The conditional expectation of the Gumbel-distributed idiosyncratic shock is a known function of the choice probability $m_{i,js}$, ensuring that W is well-defined and strictly concave in the relevant allocation variables. The planner chooses labor allocations $\{L_{i,js}\}$ (equivalently, migration shares $\{m_{i,js}\}$), capital allocations $\{K_{i,s}\}$, technology parameters $\{\rho_i\}$, sectoral consumption $\{C_{i,s}\}$, housing consumption $\{H_i\}$, and lump-sum transfers $\{T_i\}$. The planner is subject to the following constraints.

Resource constraints. Labor in each region cannot exceed the sum of migrants: $\sum_s L_{j,s} \le \sum_{i,s} m_{i,js}\,\overline{L}_i$. Total capital across all regions cannot exceed the aggregate endowment: $\sum_{i,s} K_{i,s} \le \sum_i \overline{K}_i$. The fixed factor satisfies $H_i \le \overline{\mathcal{H}}_i$, with equality in equilibrium.

Technology constraints. Output in each region-sector is produced via the CES function $Y_{i,s} = Z_{i,s}[\gamma(\rho_i)K_{i,s}^{\rho_i} + (1-\gamma(\rho_i))L_{i,s}^{\rho_i}]^{1/\rho_i}$. Sectoral composites are formed through the Armington aggregator with trade costs $\tau_{ij,s}$. The planner must respect the same CES preference structure as consumers in the decentralized economy, which means that sectoral composites are formed using the same substitution elasticity σ_s and that the price index takes the same CES form.

Institutional frictions. Workers moved from origin i to destination j incur the utility penalty μ_{ij}, exactly as in the decentralized equilibrium. Capital allocated to region i is subject to the resource cost implied by τ_i^K: to provide one unit of productive capital to firms in region i, the planner must divert τ_i^K units of capital from the aggregate endowment, with the difference $(\tau_i^K - 1)$ absorbed as a real resource cost. This formulation respects the same physical trade-off that private agents face.

Transfers. The planner can make unrestricted lump-sum transfers across workers, equalizing the marginal social utility of income. Distributional considerations therefore do not constrain the planners allocation of productive resources. The planner can also redistribute the operating profits that arise from monopolistic competition, as well as the intermediary profits Π^F.

We now derive the planners first-order conditions and compare them with those of the competitive equilibrium.

Technology choice. The planner minimizes the resource cost of producing the output vector $\{Y_{i,s}\}$. For given factor allocations and factor prices (determined endogenously by factor market clearing), the planner selects ρ_i to minimize unit cost in each region. Because the unit cost function $\mathcal{C}(\omega_i;\rho)$ is identical in the planners and the decentralized problems-it depends only on the production technology and factor prices, not on the market structure-the planners optimality

condition is $\partial \ln \mathcal{C}(\omega_i; \rho)/\partial \rho = 0$. This is precisely equation (7) of the decentralized equilibrium.

Factor allocation. Within each region, the planner allocates labor across sectors to equate the marginal value product of labor, yielding a common wage w_i. Across regions, the planner allocates capital such that the wedge-adjusted marginal value product of capital is equalized: $r_i/\tau_i^K = \overline{r}$ for some shadow price $\overline{r}$. These are identical to the factor demand equation (4) and the capital market clearing condition.

Goods markets and pricing. The planner delivers sectoral composites to consumers using the same Armington aggregator as in the decentralized economy. Because consumers have CES preferences with substitution elasticity σ_s, the planners efficient allocation of consumption across varieties yields the same CES demand structure as in the decentralized equilibrium. Moreover, the planners shadow price for each variety-the Lagrangian multiplier on the resource constraint for that variety-satisfies the same constant-markup relationship $p_{i,s} = [\sigma_s/(\sigma_s - 1)]c_{i,s}^*$ that emerges from profit maximization in the decentralized economy. This equivalence holds because the CES structure implies that the marginal rate of substitution between varieties depends only on their relative quantities and the substitution elasticity, regardless of whether quantities are chosen by a planner or by decentralized consumers. The planner cannot eliminate the markup without eliminating the product differentiation that gives rise to it under CES preferences. Consequently, operating profits $\pi_{i,s} = (p_{i,s} - c_{i,s}^*)Y_{i,s}$ are identical in the two allocations, and the planner simply redistributes these profits as lump-sum transfers.

Migration. The planner chooses migration shares $m_{i,js}$ to maximize W. The first-order condition yields the logit expression (1), because the Gumbel distribution of preference shocks and the planners objective imply that the optimal

allocation of workers across destinations follows the same log-sum formula as the decentralized choice probabilities. The planners ability to equalize the marginal social utility of income via lump-sum transfers-including the redistribution of operating profits and intermediary profits-is essential: without transfers, income differences across workers would distort migration patterns away from the decentralized outcome. With optimal transfers, the planners migration condition coincides with the private optimality condition.

Since all the planners first-order conditions are identical to those of the competitive equilibrium, and since both problems are strictly concave under the maintained assumptions, the unique solution to the planners problem coincides with the equilibrium allocation. This establishes constrained efficiency.

The critical insight is that the constant within-sector markup under CES monopolistic competition does not create a wedge between the planners allocation and the market outcome. Both the planner and private agents face the same CES preference structure, which implies the same constant-markup pricing rule. The planner cannot improve welfare by deviating from this rule, because doing so would violate the optimality conditions for consumption allocation under CES preferences. The endogenous dispersion of technologies across regions is therefore an integral part of the constrained-efficient allocation under both the capital market friction τ_i^K and the product differentiation inherent in CES preferences.

We emphasize that this result distinguishes clearly between two types of markup variation. Sector-level constant markups $\mu_s = \sigma_s/(\sigma_s - 1)$ are a feature of product differentiation and are present in both the planners and the market allocations; they do not represent misallocation. Firm-level deviations from the sectoral mean markup-which the DLW methodology is designed to detect-would, if present, constitute genuine misallocation in our framework, as they would drive a wedge between the planners and the markets allocation of factors across firms

within a sector. The constant-markup structure of our baseline model cleanly isolates the technology choice channel from this within-sector markup dispersion channel.

Finally, we note the precise sense in which the frictions are "constraints" rather than "distortions" from the planners perspective. The planner would prefer a world with $\mu_{ij} = 1$ and $\tau_i^K = 1$. The point of Proposition 3 is not that the frictions are desirable, but rather that given their existence-whether for political, institutional, or technological reasons that lie outside the model-the market achieves the best possible allocation. Any policy intervention that respects the same frictions cannot improve welfare, and any intervention that attempts to circumvent them (for example, by mandating uniform technologies) will reduce welfare. The welfare-improving reform is to reduce the frictions themselves, not to manage their symptoms through second-best interventions in technology or factor allocation.

Appendix F: Calibration Details and Additional Quantitative Results

This appendix provides supplementary details on the data construction, the estimation of $\gamma(\rho)$ under monopolistic competition, and additional quantitative results referenced in Section 9.

Data construction. The Annual Survey of Industrial Firms (ASIF) data for 1998-2013 are cleaned following the standard procedures in the misallocation literature. We exclude observations with missing or negative values for gross output, value added, employment, capital stock, or intermediate inputs. We drop firms with fewer than eight employees and trim the top and bottom 1% of the TFPR distribution within each industry-year. Nominal variables are deflated using industry-specific output and input price deflators from the National Bureau of Statistics. Real capital stocks are constructed by the perpetual inventory method with a depreciation rate of 9% per annum. The user cost of capital is computed as $r_i = (R_i + \delta)p_{K,i}$, where R_i is the province-specific lending rate from the Peoples Bank of China, $\delta = 0.09$ is the depreciation rate, and $p_{K,i}$ is the provincial investment goods price deflator. The wedges τ_i^K are then computed as $\tau_i^K = r_i/\overline{r}$, where the national benchmark rate $\overline{r}$ is the average lending rate across all provinces in each year.

Estimation of $\gamma(\rho)$ under monopolistic competition. Because the model now adopts monopolistic competition as its baseline market structure, the estimation of the technology menu must account for the wedge between revenue shares and true cost shares created by markups. The estimation proceeds in two stages. In the first stage, we apply the De Loecker and Warzynski (2012) control function approach at the firm level to estimate firm-specific markups $\hat{\mu}_i$. Using a translog production function for each two-digit industry, we estimate the output elasticity of materials θ_i^M via the Ackerberg, Caves, and Frazer (2015) procedure, and compute $\hat{\mu}_i = \hat{\theta}_i^M / s_i^M$. The markup-corrected capital share is then $\hat{s}_{K,i} = s_{K,i}^R \cdot \hat{\mu}_i$. These corrected

shares are aggregated to the province-sector level using revenue weights. In the second stage, we solve the system of the first-order condition $\partial \ln \mathcal{C}/\partial \rho = 0$ and the factor share equation

$$s_{K,i} = \frac{\gamma(\rho^*)^{\frac{1}{1-\rho^*}}}{\gamma(\rho^*)^{\frac{1}{1-\rho^*}} + (1-\gamma(\rho^*))^{\frac{1}{1-\rho^*}} \omega_i^{\frac{\rho^*}{1-\rho^*}}}$$

numerically for each province-sector observation, using the corrected shares. This yields a locus of points $(\hat{\rho}_i^*, \hat{\gamma}_i)$. The linear specification $\gamma(\rho) = \theta_0 + \theta_1 \rho$ is fitted by non-linear least squares. Standard errors are clustered at the province level to account for within-province correlation in the error term. The resulting estimates are $\hat{\theta}_0 = 0.28$ (clustered standard error 0.04) and $\hat{\theta}_1 = 0.29$ (clustered standard error 0.06), both significant at the 1% level. The estimate of $\hat{\theta}_1 = 0.29$ is slightly smaller than the value of 0.34 that would be obtained using uncorrected capital shares, confirming that markups do introduce a modest upward bias in the uncorrected estimate.

Verification of the interior solution condition. For each province-sector observation in our sample, we compute the left-hand side and right-hand side of the sufficient condition derived in Appendix A, using the markup-corrected parameter estimates. The ratio of the left-hand side to the right-hand side has a minimum value of 2.7, a median of 8.3, and a maximum of 42.1 across all observations. These values confirm that the interior solution condition is comfortably satisfied for every data point in our estimation sample.

Sectoral heterogeneity. We re-estimate the model separately for each of the three largest manufacturing sectors (textiles, machinery, and electronics). The endogenous heterogeneity shares are 35%, 42%, and 39% respectively, indicating that the mechanism is not driven by any single sector.

Inland versus coastal provinces. We estimate the technology menu using only inland provinces. The endogenous heterogeneity share is 42%, compared to 36% for coastal provinces alone, consistent with the FDI-interacted results reported in Appendix D.

WTO accession period. We estimate the model excluding the 2001-2003 period, which corresponds to Chinas WTO accession and associated tariff reductions, and obtain an endogenous heterogeneity share of 44%. This suggests that the trade liberalization episode somewhat compressed the within-sample variation in factor prices, and our baseline estimate may be conservative.

Migration elasticity. We vary the migration elasticity κ from 1.5 to 3.5. The endogenous heterogeneity share ranges from 35% to 45%, with the share increasing in κ. Higher migration responsiveness amplifies the spatial dispersion of factor prices by concentrating labor in high-wage regions, thereby widening technology differences.

Sensitivity to sectoral trade elasticities. We vary the sectoral trade elasticities σ_s within the ranges reported by Broda and Weinstein (2006). The endogenous heterogeneity share varies by less than three percentage points across the full range of plausible values, confirming that the constant-markup structure of the model does not materially affect the quantitative decomposition.

Uniform-technology counterfactual. We simulate an equilibrium in which ρ_i is forced to equal the national average $\overline{\rho}$ in all regions, while all frictions remain in place. The equilibrium is recomputed holding technology fixed and allowing factor allocations, prices, and markups to adjust until all markets clear. Aggregate real output falls by 3.8% relative to the decentralized equilibrium. The output loss is unevenly distributed: inland provinces experience a decline of 5.2% on average, while coastal provinces experience a decline of 2.1%. This pattern reflects the fact that inland provinces, where labor is abundant and capital is expensive, are forced to adopt more capital-intensive technologies poorly suited to their factor

endowments. Coastal provinces, which are closer to the national average technology, suffer a smaller loss. This counterfactual provides a direct quantitative illustration of the productive value of endogenous technology dispersion.

Hukou rent attenuation. In the baseline model, we assume that land rents are fully transferable to migrants regardless of their duration of absence from their hukou location. In reality, migrants who remain outside their home province for extended periods may see their claims to village land erode or become difficult to enforce. To assess the sensitivity of our results to this assumption, we introduce a distance-dependent attenuation factor $\delta_{ij} \in [0,1]$ such that the effective land rent received by a migrant from i living in $j \neq i$ is $\delta_{ij}\Pi_i/\overline{L}_i$. We parameterize δ_{ij} as a declining function of bilateral distance, with $\delta_{ij} = \exp(-\eta \cdot \text{dist}_{ij})$ and η calibrated to match survey evidence on the relationship between migration duration and land claim retention. Under this specification, the home bias is stronger, spatial factor price dispersion increases, and the share of TFPR variance attributable to endogenous technology heterogeneity rises to 43%. This confirms that our baseline assumption of full rent portability is conservative for the magnitude of the mechanism we identify.

Appendix G. Derivation of Firm-Level Distortions and the Structural Decomposition of TFPR

This appendix provides the detailed derivation of the firm-specific distortions from the first-order condition, the construction of the structural TFPR decomposition, and a proof that the normalization of distortion means does not affect the variance decomposition.

G.1 Recovering Firm-Level Distortions

Recall from equation (12) in the main text that the optimal capital-labor ratio for firm f located in province i is

$$\frac{K_f}{L_f} = \frac{\hat{\gamma}_i}{1-\hat{\gamma}_i}\left[\frac{w_i(1+\tau_f^L)}{r_i(1+\tau_f^K)}\right]^{\frac{1}{1-\hat{\rho}_i^*}},$$

where $\hat{\rho}_i^*$ and $\hat{\gamma}_i = \hat{\gamma}(\hat{\rho}_i^*)$ are the estimated optimal substitution elasticity and capital distribution parameter for region i, and w_i, r_i are the local wage and rental rate of capital. Taking logs and rearranging yields

$$\ln\left(\frac{K_f}{L_f}\right) = \frac{1}{1-\hat{\rho}_i^*}\left[\ln\left(\frac{\hat{\gamma}_i}{1-\hat{\gamma}_i}\right) + \ln\left(\frac{w_i}{r_i}\right)\right] + \frac{1}{1-\hat{\rho}_i^*}[\ln(1+\tau_f^L) - \ln(1+\tau_f^K)].$$

All terms on the right-hand side are known from data and prior estimation, except for the distortion pair. Let

$$\delta_f \equiv \ln(1+\tau_f^L) - \ln(1+\tau_f^K).$$

Then δ_f is identified directly from (G.2). To separate the absolute levels of τ_f^K and τ_f^L, we follow the standard procedure in the misallocation literature (Hsieh and Klenow, 2009) and normalize the means of $\ln(1+\tau_f^K)$ and $\ln(1+\tau_f^L)$ to zero within each province-sector cell. Specifically, for each (i,s) we set

$$\frac{1}{N_{is}}\sum_{f\in(i,s)} \ln(1+\tau_f^K) = 0, \qquad \frac{1}{N_{is}}\sum_{f\in(i,s)} \ln(1+\tau_f^L) = 0,$$

which together with the observed δ_f uniquely pins down the two distortion measures. This normalization implies that, on average, firm-specific distortions within a region-sector do not bias the average cost of capital or labor away from the local market prices (r_i, w_i).

G.2 Structural Decomposition of Firm-Level Log TFPR

In the monopolistic competition environment of Section 6, firm f charges a price $P_f = \mu_s \cdot c_f$, where $\mu_s \equiv \sigma_s/(\sigma_s - 1)$ is the sector-specific constant markup and c_f is the unit cost of production. For the CES production function (11) with optimal technology $\hat{\rho}_i^*$, the unit cost function can be written as

$$c_f = \frac{1}{Z_f} \cdot r_i(1+\tau_f^K)^{\hat{\gamma}_i}[w_i(1+\tau_f^L)]^{1-\hat{\gamma}_i} \cdot \mathcal{C}(\omega_i(1+\tau_f^L)/(1+\tau_f^K); \hat{\rho}_i^*),$$

where $\mathcal{C}(\cdot)$ is the minimized unit cost component derived in Section 5.2, and $\omega_i = w_i/r_i$. Note that the cost function depends on the distortion-adjusted factor price ratio. However, because the firm chooses its technology ρ_i^* based on the *local market prices* (w_i, r_i) rather than its own distortions-an assumption justified by the fact that technology choice is a longer-run decision while distortions may be firm-specific and possibly transient-the term $\mathcal{C}$ is evaluated at the undistorted ω_i (see the discussion preceding Lemma 1). This yields the exact decomposition of revenue TFP:

$$\text{TFPR}_f \equiv \frac{P_f Y_f}{K_f^{\tilde{\gamma}} L_f^{1-\tilde{\gamma}}} = \mu_s \cdot \frac{r_i^{\tilde{\gamma}} w_i^{1-\tilde{\gamma}}}{\mathcal{C}(\omega_i; \hat{\rho}_i^*)} \cdot Z_f \cdot \underbrace{\frac{(1+\tau_f^K)^{\hat{\gamma}_i - \tilde{\gamma}}(1+\tau_f^L)^{1-\hat{\gamma}_i-(1-\tilde{\gamma})}}{\text{distortion adjustment}}}_{\equiv D_f}.$$

Taking logs,

$$\ln\text{TFPR}_f = \ln\mu_s - \underbrace{\ln\mathcal{C}(\omega_i; \hat{\rho}_i^*)}_{\equiv \ln\Xi(\omega_i;\hat{\rho}_i^*)} + \tilde{\gamma}\ln r_i + (1-\tilde{\gamma})\ln w_i + \ln Z_f + \ln D_f.$$

The term $\ln\Xi(\omega_i; \hat{\rho}_i^*)$ collects all components that depend only on regional variables and the optimal technology, and is therefore common to all firms in the

same province. If we use the region's own capital elasticity $\tilde{\gamma} = \hat{\gamma}_i$, the distortion term simplifies to $D_f = (1 + \tau_f^K)^0(1 + \tau_f^L)^0 = 1$ only when both distortions are zero; otherwise, D_f captures the deviation of the firm's effective factor cost shares from those implied by the local technology. In the baseline we set $\tilde{\gamma} = \bar{\gamma}$, the national average, to maintain direct comparability with HK. Then $\ln D_f$ absorbs both the purely distortive effects and a portion of the technological mismatch between the national benchmark elasticity and the local one.

Equation (G.3) is the structural counterpart of equation (13) in the text, with $F(\tau_f^K, \tau_f^L; \hat{\rho}_i^*) \equiv \ln D_f$.

G.3 Invariance of the Variance Decomposition to Distortion Normalization

The variance decomposition in Section 9.5 separates $Var(\ln \mathrm{TFPR}_f)$ into a between-group component $Var(\ln \overline{\mathrm{TFPR}}_{is})$ and a within-group component $\mathbb{E}[Var(\ln \mathrm{TFPR}_f | i, s)]$. The endogenous technology choice component is measured by $Var[\ln \Xi(\omega_i; \hat{\rho}_i^*)]$.

We now show that the normalization $\mathbb{E}[\ln(1 + \tau_f^K)|i, s] = \mathbb{E}[\ln(1 + \tau_f^L)|i, s] = 0$ does not affect this decomposition. Let $\tilde{\tau}_f^K$ and $\tilde{\tau}_f^L$ denote any alternative normalization that differs by additive province-sector constants c_{is}^K and c_{is}^L from our baseline. Then $\ln D_f$ changes by a term Δ_{is} that depends only on (i, s). Because this term is constant within each (i, s) cell, it is fully absorbed into the cell mean $\ln \overline{\mathrm{TFPR}}_{is}$ and leaves the within-cell variance unchanged. The between-group variance changes by $Var(\Delta_{is})$, but the structural term $\ln \Xi$ is invariant to the normalization. In our procedure we define the residual regional misallocation as the difference between the total between-group variance and $Var(\ln \Xi)$; any Δ_{is} merely shifts the split between the component "residual regional misallocation" and the between-group mean, without altering the share attributed to endogenous technology choice. Thus, the central quantitative result-the fraction of TFPR

dispersion due to technology choice-is independent of the particular normalization chosen for the absolute levels of firm distortions.

Appendix H. Numerical Procedures and Moment Conditions for Technology Menu Estimation

This appendix details the numerical implementation of the technology menu estimation described in Section 9.2.

H.1 Solving for the Implicit Optimal ρ

For a given province i with observed wage-rental ratio ω_i and a candidate parameter vector $\theta = (\theta_0, \theta_1)$, the optimal substitution elasticity ρ^* is the unique solution in the admissible interval $[\underline{\rho}, \overline{\rho}]$ to the first-order condition (equation (8))

$$\frac{\partial \ln \mathcal{C}(\omega_i; \rho)}{\partial \rho} = 0, \qquad \text{where} \qquad \mathcal{C}(\omega; \rho) = [\gamma(\rho)^{\frac{1}{1-\rho}} + (1 - \gamma(\rho))^{\frac{1}{1-\rho}} \omega^{-\frac{\rho}{1-\rho}}]^{-\frac{1-\rho}{\rho}},$$

and $\gamma(\rho) = \theta_0 + \theta_1 \rho$. The derivative $\partial \ln \mathcal{C} / \partial \rho$ is computed analytically using the expressions provided in Appendix A. Because the second-order condition $\partial^2 \ln \mathcal{C} / \partial \rho^2 > 0$ holds for all empirically relevant ω under the estimated θ (the sufficient condition of Appendix A is satisfied by a factor of at least 2.7), the solution is unique. We locate it using Brent's method with a tolerance of 10^{-10}.

H.2 Moment Conditions and Objective Function

Given the solved $\rho_g(\theta)$ for observation $g = (i, s)$, the model-implied capital share is

$$s_K(\rho_g(\theta), \omega_i; \theta) = \frac{\gamma(\rho_g(\theta); \theta)^{\frac{1}{1-\rho_g(\theta)}}}{\gamma(\rho_g(\theta); \theta)^{\frac{1}{1-\rho_g(\theta)}} + (1 - \gamma(\rho_g(\theta); \theta))^{\frac{1}{1-\rho_g(\theta)}} \omega_i^{\frac{\rho_g(\theta)}{1-\rho_g(\theta)}}}.$$

The parameter vector θ is estimated by minimizing the sum of squared deviations between the markup-corrected observed capital shares $\hat{\gamma}_g$ and the model-implied shares:

$$\hat{\theta} = \arg\min_{\theta_0,\theta_1} \sum_{g=1}^{G} [\hat{\gamma}_g - s_K(\rho_g(\theta), \omega_i; \theta)]^2.$$

The minimization is performed using a quasi-Newton algorithm with analytical gradients computed via the implicit function theorem to account for the dependence of ρ_g on θ through the first-order condition. Because the objective function is smooth and locally convex near the optimum, convergence is fast and robust to starting values in the plausible range ($\theta_0 \in (0.2,0.4)$, $\theta_1 \in (0.2,0.5)$).

H.3 Standard Errors and Inference

We report province-clustered standard errors to allow for arbitrary within-province correlation of the error term $\hat{\gamma}_g - s_K(\cdot)$. The asymptotic variance-covariance matrix of $\hat{\theta}$ is estimated using the standard sandwich formula

$$\widehat{Var}(\hat{\theta}) = (G'G)^{-1} \left(\sum_{c=1}^{N_{\text{prov}}} G_{c'} \hat{\epsilon}_c \hat{\epsilon}_{c'} G_c \right) (G'G)^{-1},$$

where G is the $G \times 2$ matrix of gradients of the fitted shares with respect to θ, G_c is the submatrix for province c, and $\hat{\epsilon}_c$ is the vector of residuals within province c. In addition, we perform a nonparametric bootstrap at the province level with 1,000 replications to validate the asymptotic approximation; the bootstrap confidence intervals are nearly identical to those based on the clustered standard errors.

H.4 Verification of the Interior Solution Condition

For every observation g, we compute the left- and right-hand sides of the sufficient condition derived in Appendix A. The condition is

$$\frac{\gamma'(\rho)}{\gamma(\rho)} > \frac{1}{(1-\rho)^2} \left[1 + \omega^{\frac{\rho}{1-\rho}} \left(\frac{\gamma}{1-\gamma} \right)^{\frac{1}{1-\rho}} \right]^{-1}.$$

Using the estimated $\hat{\theta}$, the minimum, median, and maximum ratios of the left- to the right-hand side across all observations are 2.7, 8.3, and 42.1, respectively.

Thus, the condition is comfortably satisfied for the entire sample, ensuring global validity of Lemma 1.

Appendix I. Algorithm for General Equilibrium Counterfactual Simulations

This appendix describes the nested fixed-point algorithm used to compute the counterfactual equilibria in Section 9.6. The model's general equilibrium is defined in Section 7, and the counterfactuals involve changing some or all of the frictions $\{\mu_{ij}, \tau_i^K\}$ and the firm-level distortions $\{\tau_f^K, \tau_f^L\}$.

I.1 Structure of the Equilibrium

For a given set of spatial frictions and firm-level wedges, the equilibrium is characterized by the following system:

- **Technology choice:** $\rho_i^* = \underset{\rho}{\operatorname{argmin}} \mathcal{C}(\omega_i; \rho)$ for each province i, where $\omega_i = w_i/r_i$.
- **Factor prices:** $r_i = \tau_i^K \overline{r}$, with the national benchmark $\overline{r}$ adjusting to clear the aggregate capital market $\sum_i K_i = \sum_i \overline{K_i}$; wages w_i clear each provincial labor market $L_i = \sum_s L_{is}$.
- **Firm-level factor demands:** Equations (11)-(12) in the text determine the capital and labor allocated to each firm given its distortions, and aggregating yields provincial sectoral demands K_{is}, L_{is}.
- **Goods market clearing:** Under the Armington trade structure with trade costs $\tau_{ij,s}$, mill prices $p_{is} = \frac{\sigma_s}{\sigma_s - 1} \frac{r_i c_i^*}{Z_{is}}$, where $c_i^* = \mathcal{C}(\omega_i; \rho_i^*)$. Trade shares and regional incomes are determined as described in Section 6.
- **Migration:** The share of workers from i choosing to live in j and work in sector s is given by the logit expression (1) in the main text, ensuring that the realized labor supply in each province L_j is consistent with wages, prices, and hukou costs.

I.2 Nested Fixed-Point Iteration

The high dimensionality of the system and the endogenous choice of ρ_i^* make a simultaneous solution infeasible. We therefore adopt a nested iterative procedure.

Outer loop (technology update). Start with an initial guess for the vector of regional substitution elasticities $\boldsymbol{\rho}^{(0)} = (\rho_1^{(0)}, ..., \rho_N^{(0)})$. In practice, we use the values calibrated in the baseline equilibrium.

Inner loop (factor and goods market equilibrium given technology). For fixed $\boldsymbol{\rho}$, the production functions in each province are parametric. We solve for the market-clearing vector of wages $\mathbf{w}$ and the national interest rate $\bar{r}$ using the following steps:

Guess $\mathbf{w}$ and $\bar{r}$.

Compute $r_i = \tau_i^K \bar{r}$ and $\omega_i = w_i / r_i$.

With ρ_i fixed, the unit cost c_i^* is immediately obtained. Mill prices and trade shares follow.

Using the trade shares and the income formation process (including the distribution of land rents, capital income, and intermediary profits as specified in Section 4.3), compute total expenditure in each region and sector, and derive the demand for each region's output.

From the output demands, calculate the required K_{is} and L_{is} via the firm-level factor demand equations (12), after assigning firm-level distortions according to the scenario (e.g., setting $\tau_f^K = \tau_f^L = 0$ in Scenarios A-C).

Aggregate labor demand across sectors and compare to the labor supply given by the migration block. Update the wage vector $\mathbf{w}$ using a tâtonnement rule until provincial labor markets clear.

Check capital market clearing: adjust $\bar{r}$ upward (downward) if aggregate capital demand exceeds (falls short of) aggregate capital endowment.

The inner loop is solved to high precision (tolerance 10^{-6} on excess demands).

Outer loop convergence. Once the inner loop has converged for $\boldsymbol{\rho}^{(k)}$, we compute the new cost-minimizing technology for each province given the equilibrium ω_i from the inner loop:

$$\rho_i^{(k+1)} = \arg\min_{\rho} \mathcal{C}(\omega_i; \rho).$$

The outer loop iterates until $\max_i |\rho_i^{(k+1)} - \rho_i^{(k)}| < 10^{-6}$. In practice, convergence is monotonic and achieved within 15-25 outer iterations for all scenarios.

I.3 Budget Neutrality and Transfers

In all counterfactuals, the government (or the planner) is assumed to operate under budget neutrality. For scenarios that eliminate the capital wedge τ_i^K, the resource cost of intermediation disappears, and the aggregate capital endowment is directly allocated to firms without any wastage. The corresponding loss of intermediary profits Π^F is compensated by an increase in direct capital income to households, so that aggregate household income remains consistent with the economy's production possibilities.

Similarly, changes in tax-like instruments (e.g., in extensions with distortionary taxation) are offset by lump-sum transfers that equalize the marginal social utility of income across workers, as specified in the constrained planner's problem of Section 7. All welfare and TFP comparisons reported in the text are based on real consumption-equivalent measures that fully account for these transfers.

I.4 Validation

To validate the algorithm, we confirm that, starting from the baseline estimated parameters, the inner loop with $\boldsymbol{\rho}$ held at its baseline values reproduces the observed provincial wages, migration flows, and trade shares with an R^2 above 0.95. Additionally, when all frictions are retained and the baseline distortions are fed in, the solved equilibrium TFPR distribution matches the empirical distribution almost exactly (Kolmogorov-Smirnov statistic < 0.02). This ensures

that the counterfactual differences reported in Table 4 reflect genuine policy effects rather than approximation error.

Appendix J. Additional Robustness Results (Optional)

J.1 Alternative Production Function Estimators

We re-estimate firm-level markups and output elasticities using the methodology of Gandhi, Navarro, and Rivers (2020), which corrects for potential misspecification in the proxy variable approach. The resulting technology menu parameters are $\hat{\theta}_0 = 0.27$ and $\hat{\theta}_1 = 0.31$, and the endogenous technology choice share of TFPR variance becomes 27.5%. Scenario B TFP gain is +25.3%.

J.2 Different Aggregation of Distortions

Instead of normalizing the mean log distortions to zero, we normalize the medians to zero within each province-sector. The variance decomposition is virtually unchanged: endogenous technology choice share 26.9%.

J.3 Capital Share with Cost-Share Weights

In the baseline, we average capital shares using revenue weights. When cost-of-goods-sold weights are used, $\hat{\theta}_1 = 0.28$ and the technology choice variance share is 26.1%.

These additional checks confirm that the central quantitative finding-that roughly one-quarter of firm TFPR dispersion is attributable to endogenous technology choice and is not misallocation-is robust across a wide range of reasonable methodological choices.