Estimating Production Functions with Latent Team Structures: An Analysis of Nursing Homes

Nihal Mehta

The Pennsylvania State University
nzm5430@psu.edu

Last updated: September 15, 2024

Click here for the latest version

Abstract

In early 2024, the Centers for Medicare & Medicaid Services (CMS) introduced a nationwide minimum staffing mandate targeting specific worker types to address the inadequate provision of care in US nursing facilities. Firms would likely respond by adjusting their staffing both within and across the targeted and non-targeted groups based on worker substitution patterns and private incentives. To quantify this, we develop a disaggregated, flexible, and parsimonious model of team-based production. We use theories from personnel economics and organizational design to guide the specification, parameter restrictions, and sparsities. Addressing standard endogeneity concerns, we propose a novel penalized and shape-constrained GMM estimator. Using successful discharge rates to measure short-term care quality, we find gains among the bottom decile firms and a reduction in disparity. However, the policy inadvertently reduces quality in the upper decile and lowers the overall mean, rendering common stochastic dominance measures inadequate to assess the overall welfare impact.

JEL Codes: I18, L23, C13, C45, C51.

Keywords: organizational design, personnel economics, penalization, machine learning, shape restrictions, GMM, healthcare, nursing homes, LASSO

I thank Keisuke Hirano, Andres Aradillas-Lopez, and Patrik Guggenberger for all the encouragement and invaluable guidance. I also thank Paul Grieco, Connor Ryan, Bradley Setzler, Karl Schurter, and all the participants in the Applied Micro Brown-bag, Econometrics Workshop, and the Student Reading Groups conducted at Penn State.

1 INTRODUCTION

Economic outcomes depend on many inputs, yet most policies target only a few, constrained by feasibility, political realities, or the need to manage uncertainties. However, agents—guided by their incentives and underlying preferences (for consumers), or technology (for firms)—will adjust all of their inputs and not just the regulated ones. Thus, for reliable counterfactual predictions and welfare analysis, our models must capture the interdependencies among individual inputs in the preferences and technology. Driven by pragmatic concerns of dealing with numerous inputs, applied researchers often simplify their models by aggregating inputs into indices based on heuristics or by selecting models using an off-the-shelf statistical criterion like LASSO. Both approaches have limitations, as we will discuss later. Ideally, model design should leverage domain-specific institutional knowledge while being data-driven in areas where consensus among researchers is lacking. With this motivation, we propose a disaggregated, flexible, and parsimonious model of team-based production. Using theories from personnel economics and organizational design, we inform our model specification, parameter restrictions, and sparsities. Addressing standard endogeneity concerns, we propose a novel penalized and shape-constrained GMM estimator. We ground our approach in an important empirical setting: the nursing home industry, which we discuss next.

As of June 2022, 20% of Covid-19 related deaths in the US occurred in long-term care facilities[1]. Reports that followed identified understaffing as a key factor in the inadequate care provided. In response, on April 18, 2023, the Biden Administration issued an executive order for the development of minimum staffing mandates, which typically have three components: (1) target base of worker types, which can be specific (e.g., registered nurses) or broad (e.g., clinical nursing staff); (2) minimum threshold, expressed in hours per resident day (HPRD); and (3) consequences for non-compliance, such as a reduction in Medicaid/Medicare reimbursements.

On April 22, 2024, the Centers for Medicare & Medicaid Services (CMS) issued new minimum staffing standards. While the specifics of this policy will be described later, the key point is that it is targeted since it regulates only the clinical nursing

[1] see Chidambaram and Burns (2022).

staff comprising of registered nurses (RNs), licensed practical nurses (LPNs), and nurse aides (NAs)—just 3 of the 32 worker types that a nursing home can employ.

The policymaker wishes to align market outcomes with their social welfare objectives. In our context, welfare is defined as a function of the distribution of the quality of short-term and long-term care in the population. Clearly, it is monotonically increasing in the staffing of each worker type in each nursing home. Also, firms are clearly incentivized by this policy to increase the overall staffing of the targeted group. However, what is not clear is whether they will increase or decrease staffing of individual worker types within both the targeted and non-targeted groups. This would depend on the pattern of worker substitution in the underlying technology, as well as the private incentives of the firms. Given that two-thirds of U.S. nursing homes are for-profit, these incentives are unlikely to align with the social welfare objectives of policymakers. As a result, it is quite possible that this policy can have unintended effects on welfare once the firms have re-calibrated their staffing mix.

To isolate the effect of staffing mix on welfare, we assume a partial equilibrium framework with firms facing a fixed price for their services as well as fixed wages and rents in their cost schedule. Implicitly, we assume that the demand for services in the output market and the supply of labor and capital in the input market are flat and perfectly elastic at those fixed prices. Of course, for a comprehensive welfare analysis, we would need to allow for a more realistic functioning of the input and output markets as well as some general equilibrium effects. This is beyond the scope of this paper but represents an important direction for future research.

We now discuss some competing practices prevalent in empirical work and the pitfalls associated with them. First, it is common for applied researchers to put forth an ad hoc specification, and, in the process, drop variables or consolidate them into indices. This might yield parsimony and interpretability at the cost of introducing omitted variable bias and specification errors. Alternatively, one can opt for an off-the-shelf statistical technique like LASSO, which penalizes the absolute magnitude of coefficients to automatically select a subset of predictors. In practice, we might end up eliminating entire inputs or discarding higher-order and interaction terms solely based on a statistical criteria. This overlooks key aspects of the production process and economically meaningful relationships. Similarly, Principal Components Analysis (PCA) generates linear indices to maximize the variance explained by those indices in the original input matrix, but lacks an economic rationale. As a consequence, these methods often display poor finite sample performance, yielding results like negative elasticities, which, while statistically valid, are economically absurd. We corroborate this intuition through Monte-Carlo experiments and demonstrate the superior performance of our estimator in terms of the RMSE when estimating the production function and direct partial elasticities.

Moving onto our estimation results, we first note that although the data records 32 worker types, many of these distinctions are nominal rather than substantive. Based on their production roles (details of which will follow later), we condense them into five types— administrative, nursing, wellness, therapist, and specialist. We treat revenue as the outcome variable of the production function. The results align with our qualitative understanding of the inner workings of nursing homes. The method selects 8 out of the 31 possible teams that can be formed from these five worker types. Large teams with 4 or more types and small singleton teams are both unlikely to occur. Specialists are concentrated in a few large, high-value teams, while the administrative staff is agnostic to team-size and the most diffuse. For the counterfactual, we take the rate of successful discharges as a metric that captures short term care quality. We find that the policy causes an increase in the successful discharge rate among the bottom decile firms and a decrease in disparity. Meanwhile, this rate falls among the top decile and the overall mean also reduces. Such trade-offs render common decision criteria, such as first- and second-order stochastic dominance, inadequate to rank these distributions.

Besides the nursing home industry, our methodology can be useful in other contexts as well, such as skill formation models studied by labor economists. We shall discuss this more later. The subsequent sections of this paper are organized as follows: Section 2 describes the relevant institutional background and related literature. Section 3 sets up a model of firm behavior. Section 4 formalizes the econometric model and introduces the shape-constrained and penalized GMM estimator. A Monte-Carlo study is presented in Section 5 to evaluate its finite sample performance. Section 6 estimates the production function of nursing homes and the counterfactual staffing mix and health outcome distribution under the minimum staffing mandate. Section 7 offers concluding remarks and future extensions.

2 BACKGROUND AND RELATED LITERATURE

2.1 Institutional Details

Nursing home expenditures totaled $\$210$ billion in 2023, about 5% of total health care spending, up from 3% in 1965. Over the next decade, the U.S. population, like that of many advanced economies, will age rapidly. A report by the Population Reference Bureau projects a 75% increase in the number of Americans aged 65 and older needing nursing home care, equating to about 2.3 million residents in 2030. Nursing home expenditures are thus expected to grow at an annual rate of 5.3%. Government payers account for 75% of their revenue— 60% is from Medicaid, 15% from Medicare, with out-of-pocket expense and private insurance covering the rest. Furthermore, about 70% of nursing homes are for-profit[2] and a similar number are in urban neighborhoods. Compared to hospitals, the role of government payers and share of for-profit owners is much larger in this industry. Only 4% of nursing homes are hospital-based and 16% have any special care unit. Policymakers have long been concerned about the low quality of care at nursing homes, often attributing this issue to for-profit ownership (see Grabowski et al. (2013)).

We obtain facility-level annual data between 2017 and 2019 from publicly available CMS sources[3]. In each year we observe about $15,000$ unique skilled nursing homes, for a total of approximately $45,000$ observations. First let us describe the data we will use for production function estimation. Our outcome of interest is net revenue and the inputs are the number of hours worked by the different worker types as well as capital assets of the firm. The *Payroll Based Journal (PBJ) Nurse Staffing and Non-Nurse Staffing* datasets provide information on the hours that different worker types are employed for in each facility for each date. It was recorded only from 2017 onwards, which is why our analysis begins from that year. We end in 2019, before the Covid-19 pandemic fundamentally altered the industry. While it records 32 different worker types, many of these distinctions stem from accounting practices rather than genuine differences in their functional roles within the production pro-

[2] The remaining 30% are owned by non-profit entities (churches, charities, government).

[3] Many of the datasets we will mention have been organized and made available for research by the Long Term Care Focus research center at Brown University. See www.ltcfocus.org for details.

cess. We consolidate them into five broad production relevant types: administrative staff, clinical working staff, specialists, therapists, and wellness staff. This also makes the problem computationally feasible, since the number of teams (which, as we will show later are the building blocks of our production model) grow exponentially with the number of worker types. Other ways of constructing the worker types is discussed in Appendix D. Data on the financial performance of these facilities comes from the *Skilled Nursing Facility Cost Report*. Since non-profit facilities may have concerns besides profit, we focus only on only for-profit facilities for the counterfactual. After data cleanup, we are left with $8,144$ nursing homes for each of the three years $(2017-2019)$, yielding a balanced panel with $24,432$ observations.

To quantify the welfare, we would also need to estimate a model of health outcomes of patients in these facilities. Nursing homes typically provide two kinds of care: short and long term. We argue that short-term care quality can be reasonably measured using a one-quarter ahead rate of successful discharges, since short term residents typically stay for less than a quarter and a non-successful discharge implies either re-hospitalization or the patient passing away which are both unambiguously bad outcomes that we wish to prevent. If we were interested in long term care, we can use the proportion of patients who report a decline in their activities of daily living (ADL) as the health outcome of interest. This is because ADL captures essential life tasks like eating, bathing, and mobility, which are critical to the well-being and quality of life of long-stay patients. These outcomes depends on the staffing mix of the facilities, which we have already talked about. But, they also depends on various facility level attributes, such as its location, size, ownership structure, and various aspects of resident acuity. We obtain the quality measures that form the basis for the outcomes and resident acuity from the *Minimum Dataset (MDS)* and the Data from *Medicare Claims*. They are based on the average level of a nursing home's performance in certain areas of care for all the residents in that facility. We obtain the facility attributes from the *Provider Information* dataset. Summary statistics of the relevant variables can be found in Table 1 below.

We now summarize key behaviors of nursing facilities documented in the literature that we aim to capture in modeling their production function and counterfactual responses. On October 5, 2020, Massachusetts increased the minimum nurse staffing from 0 to 3.58 total nurse staff HPRD. A penalty of 2% reduction in quarterly

Table 1: Summary Statistics of Variables Used in Production and Health Models

Variable	Count	Mean	Std. Dev.	25%	50%	75%
Facility Level Resident Attributes						
Number of Beds	45 040	106.9	60.7	65.0	100.0	128.0
Number of Residents	44 682	79.3	49.0	47.0	71.0	98.0
Average Age (years)	44 682	79.0	7.3	75.4	80.0	83.8
% with Low CFS	39 389	38.3	12.7	30.0	37.2	45.5
% with High CFS	16 478	15.1	13.2	8.2	14.1	20.3
% Bedfast	21 914	20.9	13.9	13.5	19.8	27.1
% Female	43 830	65.5	12.1	58.7	67.0	73.8
% White	43 552	79.6	22.0	68.5	88.0	96.8
% Under 65	21 950	22.9	17.9	12.1	20.6	31.3
% Bladder Incontinent	44 140	79.1	13.2	71.4	81.0	88.9
% Bowel Incontinent	43 130	63.6	15.4	53.1	63.9	74.4
% with Alzheimer's/Dementia	13 989	50.3	15.0	40.5	50.5	60.3
% with Hypertension	44 091	76.4	10.8	71.0	78.0	83.8
% Schizophrenic/Bipolar	18 037	18.9	18.3	7.3	15.5	25.0
% Reporting Daily Pain	13 155	7.3	5.0	3.8	6.7	10.0
% Obese	37 100	29.4	8.3	23.8	28.8	34.3
% with ADL Decline	13 785	14.2	8.0	8.6	13.2	18.8
% Rehospitalized	44 850	16.9	7.1	12.6	16.5	20.9
% Successfully Discharged	44 450	55.0	18.9	45.0	58.4	68.6
Employment Hours by Worker Type (Thousands)						
Administrative Staff	46 181	11.0	8.4	6.0	9.1	13.5
Nursing Staff	46 181	101.7	68.8	56.3	88.2	129.6
Specialist	46 181	6.3	9.6	0.8	2.5	8.2
Wellness Staff	46 181	7.8	6.4	3.9	6.3	9.9
Therapist	46 181	12.6	11.7	4.8	9.7	16.8
Financial Metrics (Millions of USD)						
Gross Revenue	44 896	12.1	13.2	5.7	9.5	15.2
Operating Expense	44 896	2.0	8.4	0.2	1.2	3.1
Capital Assets	40 214	6.8	10.3	1.4	2.8	7.1

[a] A unit of observation in this sample is a (facility, year) pair.
[b] % ADL decline: long-stay residents with an increased need for help with activities of daily living.
[c] CFS is Cognitive Function Scale score. Low is 1 and high is ≥ 4 (severe cognitive impairment).

Medicaid payments was introduced from the following year for non-compliance. The treatment group was thus the subset of nursing homes with strongest incentives to respond: those with high Medicaid resident shares ($\geq$ 75th percentile) and initial staffing below the policy threshold (HPRD ≤ 3.58). Meeting these criteria are $1,617$ out of $15,333$ nursing homes nationally and 40 out of 373 nursing homes in Massachusetts. Following Abadie (2021), a CMS study uses synthetic control to construct a suitable control group for Massachusetts (optimally weighting donor states to match pre-treatment trends from 2015Q1 - 2020Q3). They found a statistically significant effect of this policy on the treatment group, and disproportionate hiring of less expensive nurse types to meet requirement. Non-targeted worker hours were also affected. Two key takeaways emerge. First, firms act in line with cost minimizing behavior and second, substitution effects between the different worker types can cause a change in the staffing of non-targeted workers as well. This is in line with what we would expect given that roughly two-thirds of all nursing homes in the US are for-profit.

In a time-use survey conducted by CMS in 2023, a group of clinicians shadowed the nursing staff for their $8-12$ hours shifts and recorded the frequency and duration of different tasks performed. The O*NET dataset, managed by the US Bureau of Labor Statistics, also floats a questionnaire to workers in which they rate themselves on the skills they possess and their ability to perform certain tasks. Both of these surveys point us towards a common qualitative view regarding the internal organisation of nursing homes: Revenue[4] is generated by the successful treatment of patients who arrive with certain ailments which requires workers to collaborate with one another. However, they do it selectively— failures in coordination and task prioritization, as is evident from delayed and omitted care, point to barriers in team formation. We use this as a foundation to develop an economic model that matches task bundles to teams of workers, incorporating costly team formation.

2.2 Related Literature

Our paper contributes to two major strands of literature. First is the body on work on the economics of nursing homes. Many papers focus on counterfactuals involv-

[4]Medicaid/Medicare reimbursement involves a complex process that takes into account various services provided to patients, which we are abstracting away from. It is a combination of direct billing for specific services and bundled payments (daily rates) that cover a range of care services.

ing Medicaid and Medicare, such as Hackmann (2019) who measures the impact of Medicaid reimbursement rates on staffing levels. Another set of papers investigate the potential misalignment of private incentives, such as Gandhi (2023) who focuses on selective admission practices, and Gupta et al. (2023) who estimate the effect of private equity infusions. A third common theme involves analyzing the welfare effect (such as the impact on patient mortality) of firm conduct as in Lin (2015) (given the oligopolistic market structure of the nursing home industry), limited supply as in Ching, Hayashi, and Wang (2015), and firm exits as in Olenski (2023). We add to this literature by studying the impact of minimum staffing policies on labor demand and the distribution of health outcomes in the population.

The second intersection is with the econometrics of latent group memberships. Bonhomme and Manresa (2015) propose grouping objects based on the K-means clustering algorithm while Su, Shi, and Phillips (2016) propose a variant of the LASSO called C-LASSO for shrinking individual parameters to an unknown group-level parameter vector. Both papers focus primarily on grouping the fixed effects in a panel data setting so as to bypass the incidental parameter problem. So, their target is different from ours (we are interested in grouping a collection of disaggregated inputs), though their techniques can be extended to other settings. For example, Almagro and Manresa (2021) apply the K-means algorithm in a BLP framework to estimate the latent structure of a nested logit model. Like Kasahara, Schrimpf, and Suzuki (2023), we also allow for the same object to be in multiple groups. But their focus is on latent firm types. They use a Gaussian mixture model and interpret the multiple memberships as probabilities. Penalties have been used for implementing various kinds of selections, such as selecting moments in Caner, Han, and Lee (2018), instruments in Liao (2013), and product attributes in a BLP setting in Gillen et al. (2019). We use penalties to select the organization design in a team based production process. Bonhomme (2021) also studies team production but their goal is to identify and estimate the individual ability of workers when we observe a collection of teams of different sizes. There is a large literature on estimation and inference under shape constraints, which we will not discuss in detail. Some important papers in this area are Menzel (2022) Matzkin (2013), Blundell, Chen, and Kristensen (2007), Wang (2023), Haag, Hoderlein, and Pendakur (2009), Chernozhukov, Newey, and Santos (2023), and Chetverikov and Wilhelm (2017).

3.1 Team Based Production Technology

As mentioned before, we will base our model on the literature that characterizes worker–job matching in a skills-tasks framework. A seminal paper here is Acemoglu and Autor (2011), who set up their model assuming a scalar skill dimension. Skills are made multi-dimensional in Lindenlaub (2017), while Ocampi (2022) allows for a continuum of tasks. We extend this framework to accommodate teams.

Suppose there are d_s number of skills, with the intensity of each being normalized to the unit interval. This yields the skill space which is a hypercube, $\mathcal{S} \equiv [0,1]^{d_s}$. Suppose $d_s = 2$, capturing two important skill dimensions in the nursing home context: mental health skill and physical health health. Next, suppose there are two worker types: nurse and wellness staff, represented by $j = 1, 2$ respectively. Generalization to an arbitrary number of workers is direct. Each type j supplies L_j hours inelastically to the firm and is fully characterized by their skill level $w_j \in \mathcal{S}$.

Next we define the team as a non-empty collection of worker types. With just two workers, the set of all teams is given by $\mathcal{G}^* = \{\{1\}, \{2\}, \{1, 2\}\}$. Here, $g = \{1\}$ and $g = \{2\}$ denote the singleton teams of the nurse and wellness staff working alone respectively, while $g = \{1, 2\}$ represents the team of both working together. The firm can split the labor endowment of the workers between the different teams, with a_{gj} denoting the proportion of hours of worker j that are allocated to team g.

In this skills based framework, we conceptualize a team as follows: consider the team $g = \{1, 2\}$, with the nurse $j = 1$ contributing $a_1 L_1$ hours and the wellness staff $j = 2$ contributing $a_2 L_2$ hours. The team pools together the resources of its constituent members, namely the skill levels and the time allocation. Suppose the pooled skill is an average of the skills of its members, so $w_g \equiv \sum_{j \in g} w_j / |g|$, while the time resources are pooled in a Cobb-Douglas fashion, denoted by $\phi_g(L; a, \gamma) = (a_{g1} L_1)^{\gamma_{g1}} (a_{g2} L_2)^{\gamma_{g2}}$, where γ_{gj} is the team-time elasticity of worker j in team g.

We now move onto defining tasks. A task t is a specific activity that needs to be done to realize the output of the firm. It can be fully characterized by its skill requirements, so $t \in \mathcal{S}$. We assume that there is a continuum of such tasks spread over the skill space with some distribution F. Given teams and tasks, we need to find a way to match them. Define the match surplus as the revenue realized when team g performs task t for 1 unit of time. Suppose it has the following specification:

$$v_{gt} = \exp\left(\underbrace{\left[w_1 t_1 + w_2 t_2\right]}_{\text{skill complementarity}} - \underbrace{\left[(t_1 - w_1)^2 + (t_2 - w_2)^2\right]}_{\text{skill mismatch}}\right) \tag{1}$$

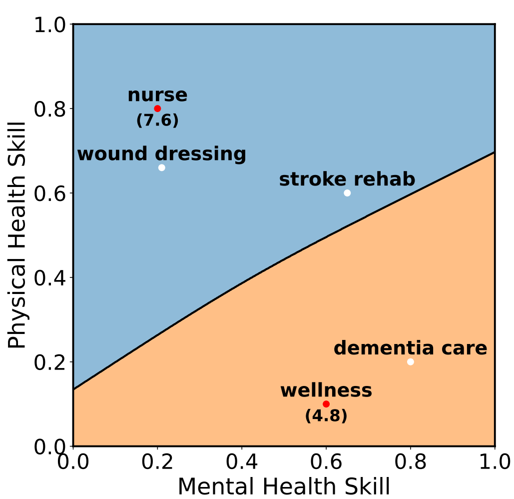
Skill complementarity captures the fact that the surplus increases with the skill requirements of the task and the skill levels of the team, but gets magnified when the task and team skills are aligned along the same dimension. Skill mismatch penalizes any deviation between the task's requirements and the team's skill level.

Next, based on the notion of comparative advantage, we assign tasks to teams so as to maximize the match surplus. The set of tasks allocated to team g is given by:

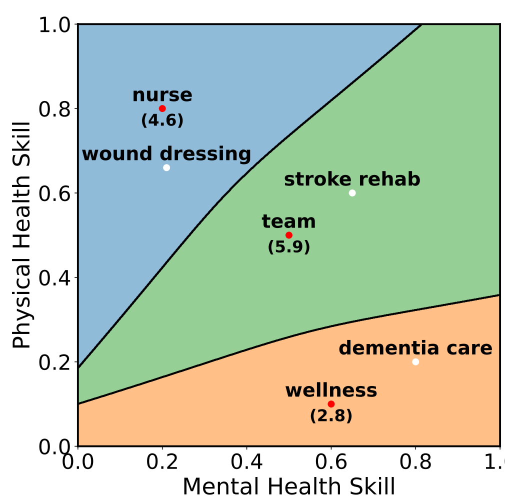
$$\mathcal{T}_g = \{t \in \mathcal{S} : v_{gt} > v_{g't} \text{ for all } g' \in \mathcal{G}^* \text{ such that } g' \neq g\}. \tag{2}$$

We do not worry about ties since the set of tasks for which that holds is of measure zero. To build intuition for all that we have discussed thus far, look at the right panel of Figure 1 below. We see two worker types: nurse and wellness staff, as well

Figure 1: Task Allocation Under Two Different Organization Designs



(a) Task Allocation when $\mathcal{G} = \{\{1\}, \{2\}\}$



(b) Task Allocation when $\mathcal{G} = \{\{1, 2\}\}$

as their team whose embedding in the skill space is determined by the skill pooling we discussed above. Tasks are uniformly distributed, with 3 of them illustrated by a white dot. Their assignment to the workers and the team is shown by different colors, with a black hyperplane separating them. Consider a task like wound dressing, which requires physical care skills, such as medical precision and attention to hygiene, making it well-suited to the nurse's expertise. In contrast, dementia care relies heavily on mental care skills, such as providing emotional stability and cognitive engagement, areas where the wellness staff excels. Stroke rehabilitation requires a balance of both. Thus, the team of both workers should collaborate to perform it, with the nurse helping the patient regain their strength and the wellness staff helping them manage their frustration. If collaboration is restricted, as in the left panel, then the nurse may take the lead due to a better skill alignment, despite neither worker being a good fit. Lastly, tasks like patient ambulation (not shown in the figure) require minimal physical and mental care. The specification in Equation 1 captures that fact that it would generate lower revenue compared to a more demanding task like stroke rehab which requires higher levels of both skills.

We can now obtain the match surplus corresponding to team g, denoted by v_g, by simply integrating v_{gt} over all tasks t that are assigned to g. Formally,

$$v_g = \int_{\mathcal{T}_g} v_{gt} \, dF(t). \tag{3}$$

The output of team g is then simply v_g scaled by the effective units of time worked by that team, $\phi_g(L; a, \gamma)$. We are now in a position to look at the firm's choice of organization design i.e. the collection of teams that it employs as well as the allocation of workers' hours within those teams. We will only look at the intuition here, with the formal notation deferred to the next subsection. In the 2 worker case, there are 8 different organization designs to choose from, with details in Table 2. Given organization design $\mathcal{G} \subseteq \mathcal{G}^*$, the revenue is the surplus generated from all the tasks done, which is simply the sum of the revenue from each of the teams:

$$\sum_{g \in \mathcal{G}} v_g \, \phi_g(L; a, \gamma).$$

Note than adding teams is beneficial to the firm since it allows better performance of some tasks leading to higher match surplus. But there is no free lunch. The firm

Table 2: Organisation Designs in Illustrative Model with Two Worker Types

S.No	Org Design $\mathcal{G}$	# Teams	# Teams (Nurse)	# Teams (Wellness)
1	{{1}, {2}}	2	1	1
2	{{1}, {1,2}}	2	2	1
3	{{2}, {1,2}}	2	1	2
4	{{1,2}}	1	1	1
5	{{1},{2},{1,2}}	3	2	2
6	{{1}}	1	1	0
7	{{2}}	1	0	1
8	{{}}	0	0	0

has to incur fixed costs whenever it attempts to instantiate any new team.

There are primarily two types of costs. First is the coordination cost which increases with the number of workers in any team. Larger teams may face communication barriers, increased potential for conflict, and difficulties in managing workflow, which can diminish overall productivity. Second is the management cost, which increases with the number of teams. More teams necessitate more managers, more meetings, and potentially more layers of bureaucracy, all of which contribute to higher administrative and overhead expenses. There is a literature on micro-foundations of these costs. For example, in Caplin et al. (2023), the costs emerge from rational inattention of the managers who decide on the worker-task assignment. On the other hand,Dessein, Lo, and Minami (2022) attributes the costs to principal-agent incentive misalignments within the organizational hierarchy.

For a given organization design $\mathcal{G}$, the firm chooses an allocation vector which maximizes the total revenue net of costs associated with that design. This optimization is subject to the allocation vector satisfying two constraints. First, it must be non-negative and second, it must respect the time budget constraint, i.e., it should sum up to exactly unity for each worker across the different teams. Designs 6, 7, and 8 in Table 2 are automatically rejected since they fail to meet this constraint.

Among the five remaining designs, consider the extremes: designs 1 and 4. With high management costs, the firm would consolidate and choose design 4— a single large team with both worker types. Conversely, under high coordination costs, the firm would decentralize and opt for design 1— two singleton teams working independently. Regardless of the specific cost structure, the key takeaway is that the firm is likely to employ only a subset of all possible team configurations.

Our model and motivation are closely related to the divisional structure explored in personnel economics as in Lazear and Gibbs (2014) and organizational design as in Burton, Obel, and Håkonsson (2021). It builds on the principle that inputs within a group are complementary in generating the group's output, while outputs across groups are substitutable in contributing to the firm's overall output. In our framework, complementarity among workers within a team is captures by the resource pooling, whereas the substitutability between teams is captured by aggregating the effective surplus generated by each team to arrive at the total revenue.

Besides the divisional structure, another popular class of organization designs is referred to as the functional structure where inputs within a group are substitutable, while outputs across groups are complementary in producing the overall output. This can be relevant for production in other contexts. For example, suppose you are a labor economist interested in skill formation. You would like to know whether a remediation policy that targets parental monetary investment can narrow the ability gaps across children. It could be that inputs such as the school's spending on learning resources and teacher's instructional time are substitutable in determining 'school quality' while the time that the parents play with their child, and the family income determine 'house quality'. The production of the final output i.e. test scores can then be modelled as being complementary in these two qualities. It could also be that the school's expenditures and family income are substitutable in determining 'monetary investment' while the instructional and parental times determine 'temporal investment'. The test scores can then be complementary in these two investments. In such a case, our methodology offers researchers the ability to incorporate their domain knowledge of this functional structure while letting the grouping choice remain data-driven.

3.2 Firm's Choice of Organization Design

We first extend the previous model to a finite but arbitrary number of worker types $\{1, \ldots, d\}$. Each type j supplies L_j unit of time inelastically, with $L = (L_1, \ldots, L_d)'$.

A team is defined as a non-empty collection of worker types. We denote the set of all possible teams by $\mathcal{G}^* \equiv \mathcal{P}(\{1,\ldots,d\})\backslash\{\emptyset\}$, with $\mathcal{P}(\cdot)$ being the power set. Let a_{gj} represent the proportion of hours of worker j allocated to team g. Each team g performs a fixed bundle of tasks for an effective time given by $\phi_g(L; a, \gamma)$ as argued previously. Here γ_{gj} is the team-time elasticity of worker j in team g. Let v_g be the match surplus of this team. Then, the revenue from team g is given by:

$$v_g \phi_g(L; a, \gamma) \equiv v_g \prod_{j \in g} \left[a_{gj} L_j\right]^{\gamma_{gj}} . \tag{4}$$

A more general Nested CES form of pooling workers' time within teams and aggregating team outputs is described in Appendix B. Next, we define an organization design to be a collection of teams $\mathcal{G} \subseteq \mathcal{G}^*$ along with a feasible allocation matrix $A_{\mathcal{G}}$ that dictates how much time each worker type devotes to each team. Let $a_g = a'_{g\cdot}$ be the allocation vector for team g and let $a = \left(a_g \text{ for } g \in \mathcal{G}^*\right)$ be the stacked allocation vector across all teams. Then, the set of feasible allocation vectors $A_{\mathcal{G}}$ is:

$$A_{\mathcal{G}} = \left\{ a \in \mathbb{R}^{d2^{d-1}} \;\middle|\; a_{gj} = 0 \text{ if } g \notin \mathcal{G} \text{ else } a_{gj} > 0, \left\|a_{\cdot j}\right\|_1 = 1 \right\} . \tag{5}$$

For example, in the case with 2 workers that we described in the previous subsection, consider the design $\mathcal{G} = \{\{1\},\{1,2\}\}$. Then a_{22} is equal to zero to meet the time budget constraint and the space of feasible allocations $A_{\mathcal{G}}$ is simply given by:

$$A_{\mathcal{G}} = \left\{ a \equiv \left[a_{11}, 0, 1 - a_{11}, 1\right]' \in \mathbb{R}^4 \;\middle|\; a_{11} \in (0,1) \right\} . \tag{6}$$

Next we define the costs that intuitively introduced in the previous subsection. A firm with an organizational structure $\mathcal{G}$ incurs two types of costs: The coordination cost, which increases with the number of workers in any team $\sum_{g \in \mathcal{G}} |g|$, and the management cost, which increases with the number of teams $|\mathcal{G}|$.

The firm's objective is to choose the optimal organizational structure and allocation of labor hours to maximize total revenue while minimizing coordination and management costs. The optimization problem can be formulated as follows:

$$\max_{\mathcal{G} \subseteq \mathcal{G}^*, a \in A_{\mathcal{G}}} \sum_{g \in \mathcal{G}} v_g \prod_{j \in g} \left[a_{gj} L_j\right]^{\gamma_{gj}} - c\left(\sum_{g \in \mathcal{G}} |g|, |\mathcal{G}|\right) , \tag{7}$$

where $c(\cdot)$ is the cost function, assumed to be monotonically increasing in both of its arguments. Higher management cost will reduce the number of selected teams while a higher coordination cost reduces the size of the selected teams.

The allocation parameters must satisfy $a_{gj} \geq 0$ and $\left\|a_{\cdot j}\right\|_1 = 1$.The latter ensures that all the hours of each worker are accounted for, reflecting the time budget constraint. Choosing the optimal allocation boils down to solving a constrained optimized problem using the Karsh-Kuhn-Tucker (KKT) conditions. Unless a worker is part of only a single team in the selected design, a boundary allocation is never be optimal for that worker since that means that there exists a team that the firm could simply drop without affecting its revenue while it would save on the costs it had to incur to set up that team in the first place.

The team-time elasticity parameters are bounded below by some strictly positive number, so $\gamma_{gj} \geq \underline{\gamma} > 0$. This is because in the economics of occupation, we often assume that the surplus from any worker being matched to any task is strictly positive depending on the distance between the task and the worker in the skill space, as in **??**. Furthermore, we would expect each team to exhibit diminishing returns, with $\left\|\gamma_{g\cdot}\right\|_1 \leq 1$.

4 PENALIZED AND SHAPE-CONSTRAINED GMM ESTIMATOR

4.1 Econometric Model

There is a vast literature on estimating production functions. Endeogeneity is a primary concern here. There might be certain factors not be observed by the econometrician that not only affect the output of the firm directly but also its choice of inputs. We assume that all such factors can be consolidated into a scalar and Hicks-neutral catch-all term labeled productivity, which is allowed to be firm and time specific. Depending on the assumptions we are willing to make regarding input timings and the law of motion of productivity, there are two broad identification strategies. The first one is informally dubbed as the 'proxy variable' approach. The seminal paper here is by by Olley and Pakes (1996), followed up with important contributions from Levinsohn and Petrin (2003) and Ackerberg, Caves, and Frazer

(2015). The other body is work is dubbed the 'dynamic panel' approach. The primary paper here is Blundell and Bond (2000). Going forward we shall stick to the latter approach, as set up in Ackerberg (2023). The first strategy requires extending our framework to a semi-nonparametric setting, since some auxiliary functions are estimated using sieves (see Appendix A). Papers such as Valmari (2023), Demirer (2020) and Ackerberg, Hahn, and Pan (2022) focus on multi-product firms, multi dimensional productivity, and Non-Hicksian factor augmenting productivity. Extensions to those settings are possibly interesting directions for future work.

Consider a balanced panel of firms $i = 1, \ldots, n$ over periods $t = 1, \ldots, T$. A firm's output, capital, and labor are denoted by Y_{it}, K_{it}, L_{it} respectively, with their log values denoted by the corresponding lowercase letters. Y_{it} and K_{it} are scalars while labor $L_{it} \equiv (L_{1it}, \ldots, L_{dit})'$ is a d-dimensional vector, with L_{jit} being the number of hours employed of the jth worker type. Let ω_{it} and ε_{it} be scalar, Hicks neutral, firm and time specific productivity and idiosyncratic shocks respectively. The latter can also be interpreted as an exogenous measurement error in output. For each firm i and time t, the researcher observes Y_{it}, K_{it} and L_{it} but not ω_{it} and ε_{it}. Firms are sampled from an underlying population with the asymptotics of letting the number of firms $n \to \infty$ for a fixed T. So, the data takes the form of a short panel from which we can recover the joint distribution of $\{(y_{it}, k_{it}, \ell_{it})\}_{t=1}^{T}$. We assume that the different worker types enter into the production function through a labor index $H(\cdot) : \mathbb{R}^d \to \mathbb{R}$ with the following team-based specification:

$$H(L_{it}; \theta) = \sum_{g \in \mathcal{G}^*} v_g \prod_{j \in g} \left[a_{gj} L_j \right]^{\gamma_{gj}}, \tag{8}$$

where $\theta_g = (a'_{g\bullet}, \gamma'_{g\bullet}, v_g)'$ collect all parameters for team g, and $\theta = \big(\theta_g \text{ for } g \in \mathcal{G}^*\big)$. Equation 8 is similar to the benefit in Equation 7, except that we allow for all possible teams since a priori, the researcher does not know the true organization design. Let β_k be the output elasticity of capital. Assuming Cobb-Douglas technology between the labor index and capital, the value-added production function becomes:

$$y_{it} = \delta_0 + h(L_{it}; \theta) + \beta_k k_{it} + \omega_{it} + \varepsilon_{it}, \quad \text{where } h(L_{it}; \theta) \equiv \log H(L_{it}; \theta). \tag{9}$$

Observe that there is no i subscript in the labor index. In other words, we are assuming that all the nursing homes face the same management and coordination cost structure, and thus the organization design is common in the whole industry. This may or may not be a strong assumption. Although allowing for latent firm types is a big extension, we can allow the production technology to differ based on known attributes of the firms. For example, if we except urban nursing homes to have a systematically different organization design compared to their rural counterparts, then we can simply estimate the production functions separately for these two subsets of firms. This is a robustness exercise we are currently undertaking. Next, we make the following assumptions on input timings and latent variables:

Assumption 4.1 (Information Set and Latent Variables)**.**

1. The firm's information set just before realizing the idiosyncratic shock ε_{it}:

$$\mathcal{I}_{it} \equiv \{k_{is}, \ell_{is}, \omega_{is}, y_{is-1}\}_{s=1}^{t}, \quad \text{for } i = 1, \ldots, n, \text{ and } t = 1, \ldots, T.$$

2. Given innovation ξ, ω follows a first order auto-regressive AR(1) process:

$$\omega_{it+1} = \delta_1 \omega_{it} + \xi_{it+1}, \quad \text{for } i = 1, \ldots, n, \text{ and } t = 1, \ldots, T-1.$$

3. The innovation shock ξ has mean zero conditional on past information:

$$\mathbb{E}\left[\xi_{it+1} \mid \mathcal{I}_{it}\right] = 0, \quad \text{for } i = 1, \ldots, n, \text{ and } t = 1, \ldots, T-1.$$

4. The idiosyncratic shock ε is mean zero conditional on current information:

$$\mathbb{E}\left[\varepsilon_{it} \mid \mathcal{I}_{it}\right] = 0, \quad \text{for } i = 1, \ldots, n, \text{ and } t = 1, \ldots, T.$$

Note that the parameters of interest are $\alpha_0 \equiv (\theta_0, \beta_0) \in \Theta \times \mathcal{B} \equiv \mathcal{A}$, where θ_0 characterizes the labor index and $\beta_0 \equiv (\beta_k, \delta_0, \delta_1) \in \mathbb{R}^3$ is a finite dimensional vector. Let $X_{it} \equiv (y_{it+1}, y_{it}, k_{it+1}, k_{it}, L_{it+1}, L_{it})'$ and define the δ_1-differenced residual:

$$\begin{aligned}
\rho(X_{it}, \alpha_0) &= \delta_0 + y_{it+1} - h(L_{it+1}; \theta_0) - \beta_k k_{it+1} - \delta_1 \left(\delta_0 + y_{it} - h(L_{it}; \theta_0) - \beta_k k_{it}\right) \\
&= \omega_{it+1} + \varepsilon_{it+1} - \delta_1 \left(\omega_{it} + \varepsilon_{it}\right) && \text{(Equation 9)} \\
&= \delta_1 \omega_{it} + \xi_{it+1} + \varepsilon_{it+1} - \delta_1 \left(\omega_{it} + \varepsilon_{it}\right) && \text{(Assumption 4.1.2)} \\
&= \xi_{it+1} + \varepsilon_{it+1} - \delta_1 \varepsilon_{it}. && (10)
\end{aligned}$$

Under Assumption 4.1, our model yields $T-1$ conditional moment restrictions:

$$\mathbb{E}\left[\rho\left(X_{it}, \alpha_0\right) \mid \mathcal{I}_{it}\right] = 0, \quad \text{for } t = 1, \ldots, T-1. \tag{11}$$

This follows from the fact that $\mathbb{E}\left[\xi_{it+1} \mid \mathcal{I}_{it}\right] = 0$ from Assumption 4.1.3, $\mathbb{E}\left[\varepsilon_{it} \mid \mathcal{I}_{it}\right] = 0$ from Assumption 4.1.4, and $\mathbb{E}\left[\varepsilon_{it+1} \mid \mathcal{I}_{it}\right] = 0$ from the law of iterated expectations and Assumption 4.1.4. Note than the vector X_{it} in Equation 11 can be partitioned into 2 sets—the endogenous variables $\widetilde{X}_{it}$, and the pre-determined variables $\widetilde{\mathcal{I}}_{it}$, which are a subset of the information set $\mathcal{I}_{it}$. For the information vector $I_{it} \equiv \operatorname{vec}\left(\mathcal{I}_{it}\right)$, we consider transformations $\{Z_{imt}\}_{m=1}^{d_{znt}}$ which we collect into a vector $Z_{it}^{d_{znt}} \equiv \left(Z_{i1t}, \ldots, Z_{id_{znt}t}\right)'$. This is a sequence of known basis functions which can approximate any real-valued square integrable function of I_{it} arbitrarily well as $d_{znt} \to \infty$. Setting $I_i \equiv \left(I_{i1}, \ldots, I_{iT-1}\right)'$, we further set $d_{zn} \equiv \sum_{t=1}^{T-1} d_{znt}$, $Z_i^{d_{zn}} \equiv \left(\left(Z_{i1}^{d_{zn1}}\right)', \ldots, \left(Z_{iT-1}^{d_{znT-1}}\right)'\right)'$, and $\rho(X_i, \alpha) \equiv \left(\rho\left(X_{i1}, \alpha\right), \ldots, \rho\left(X_{iT-1}, \alpha\right)\right)'$. We are now ready to define the following object:

$$\psi_i^{d_{zn}}(\alpha) \equiv \underset{((T-1)\times 1)}{\rho\left(X_i, \alpha\right)} * \underset{(d_{zn}\times 1)}{Z_i^{d_{zn}}} \equiv \underset{(d_{zn}\times 1)}{\begin{bmatrix} \rho\left(X_{i1}, \alpha\right) \otimes Z_{i1}^{d_{zn1}} \\ \vdots \\ \rho\left(X_{iT-1}, \alpha\right) \otimes Z_{iT-1}^{d_{znT-1}} \end{bmatrix}}. \tag{12}$$

For each α, we take the product of each residual $\rho\left(X_{it}, \alpha\right)$ with $Z_{it}^{d_{znt}}$, a vector of transformations of its information vector I_{it}. So, we have generated d_{zn} unconditional moments from $T-1$ conditional moments. This, along with Equation 11, imply the following increasing number of unconditional moment restrictions:

$$\mathbb{E}\left[\psi_i^{d_{zn}}(\alpha)\right] = 0, \quad d_{zn} \to \infty. \tag{13}$$

In this paper, we study asymptotics at a researcher specified number of moment conditions fixed at d_z that does not grow with sample size. So, $d_{zn} = d_z$ for all n, and we drop this superscript from here on. While common in practice, this may result in inefficiency (as the estimator is not exploiting all the information in our model) and potential identification concerns. We are working on extending our framework to accommodate an increasing number of moment conditions.

The usual GMM objective for an $O_p(1)$ positive definite $d_z \times d_z$ weight matrix $W_n(\alpha)$:

$$Q_n^0(\alpha) = \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\alpha)\right]' W_n(\alpha)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\alpha)\right]. \tag{14}$$

4.2 *Shape-Constrained and Penalized GMM Estimator*

In subsection 4.1, we proposed an econometric model for production function estimation that embeds in Blundell and Bond (2000) a team-based specification for the labor index. In this labor index, there are 2^{d-1} allocation parameters for each worker type, 2^{d-1} team-time elasticity parameters for each worker type, and $2^d - 1$ match surplus parameters, one for each non-empty team. Other than that, we have three additional parameters not related to the labor index: the constant δ_0, the output elasticity of capital β_k, and the auto-regressive coefficient for productivity δ_1. Thus, the total number of parameters in the model is given by $d_\alpha = d(2^{d-1}) + d(2^{d-1}) + 2^d - 1 + 3 = (d+1)2^d + 2$. This grows exponentially with the number of worker types. We could end up in a high-dimensional regime with $d_\alpha \gg n$. Fortunately, the information we get from the model in Section 3 will help us considerably restrict our parameter space. That is our goal in this subsection.

It makes sense for a worker to contribute some strictly positive amount to the effective units of time for them to be meaningfully considered part of a team. Thus, team-time elasticities are bounded below by a small known strictly positive number $\underline{\gamma}$. Further, we require that γ_{gj} sums up to atmost unity across workers which ensures weakly diminishing returns to scale for each team, so $\gamma_{gj} \geq \underline{\gamma}$, and $\sum_{j=1}^{d} \gamma_{gj} \leq 1$. Since adding teams is always beneficial, as seen in equation -, we require the match surplus from each team to bounded below by some small but strictly positive known number $\underline{v}$. Allocation is non-negative and respects the time budget constraint, adding up to exactly unity for each worker type j, so $a_{gj} \geq 0$ and $\sum_{g=1}^{G} a_{gj} = 1$. Among other things, these constraints ensures that the overall labor index is monotonic and concave in its arguments.

Also, for the productivity process to converge to a stationary distribution, we require the auto-regressive coefficient to be less that unity in absolute value. Finally, for the output elasticity of capital to be meaningful, it has to lie in the unit interval.

Note that the parameter restrictions described thus far are simple linear equality and inequality constraints. So, the shape-constrained parameter space $\widetilde{\mathcal{A}} \subset \mathcal{A}$ is convex. From a computational perspective, it is straightforward to implement this in a constrained optimization routine. This parameter space is characterized as:

$$
\widetilde{\mathcal{A}} = \left\{ \alpha \in \mathcal{A} \equiv \mathbb{R}^{2^d(d+1)+2} : \begin{array}{ll} (1) & a_{gj} \geq 0 \text{ for each } j, \text{ and } g \in \mathcal{G}^*, \\ (2) & \left\| a_{\bullet j} \right\|_1 = 1 \text{ for each } j, \\ (3) & \gamma_{gj} \geq \underline{\gamma} > 0 \text{ for each } j, \text{ and } g \in \mathcal{G}^*, \\ (4) & \left\| \gamma_{g\bullet} \right\|_1 \leq 1 \text{ for each } g \in \mathcal{G}^*, \\ (5) & v_g \geq \underline{v} > 0 \text{ for each } g \in \mathcal{G}^*, \\ (6) & \beta_k \in [0,1], \\ (7) & \delta_1 \in [-1,1] \end{array} \right\}. \tag{15}
$$

Next, the costs associated with implementing a particular organization design as in Equation 7 promote sparsity in the collection of allocation vectors $\left\{a_{g\bullet}\right\}_{g\in\mathcal{G}^*}$ in the sense that $\sum_{g\in\mathcal{G}^*} \mathbf{1}\left\{a_{g\bullet} \neq \mathbf{0}\right\}) \ll |\mathcal{G}^*|$. Let $p_g \equiv \dim(a_{g\bullet})$ represent the number of workers in team g, and $\left\|a_{g\bullet}\right\|_2 = \sqrt{\sum_{j\in g} a_{gj}^2}$ be the ℓ_2-norm of the allocation vector. Although the exact cost structure is not known to the researcher, we know that it has two arguments—coordination and management costs—and increases monotonically in both. We propose the a penalty based on these considerations:

$$
\text{Penalty} = \lambda_n \sum_{g=1}^{G} \sqrt{p_g} \left\| a_{g\bullet} \right\|_2 . \tag{16}
$$

We now examine the components of this penalty. Firstly, coordination cost implies that organizations are unlikely to maintain large teams, which we capture in our penalty via the term $\sqrt{p_g}$—increasing the penalty at a diminishing rate as team size grows. Secondly, management cost suggests that the number of teams in an organization should remain limited, reflected in the term λ_n which we leave as a researcher-specified tuning parameter. The penalty targets the ℓ_2-norm of the allocation vectors, incentivizing the estimator to eliminate entire allocation vectors for some teams. While we could have applied this selection process to the match surplus parameters v_g, doing so would risk allocating non-zero worker hours to teams with zero match surplus, potentially invalidating the time budget constraint in Equation 15. Moreover, firms' optimizing behavior suggests that allocation in teams not part of the chosen design will be zero, whereas match surplus is a primitive we expect to remain positive, even if small, irrespective of the chosen design.

Thus, our economically motivated team sparsity is implemented using a group-LASSO type penalty. Readers interested in understanding how group-LASSO can zero out entire vectors of parameters may refer to Proposition 1 in Yuan and Lin (2006). Besides aligning with the economic objective, this penalty also serves as a statistical regularization criterion that promotes model parsimony. So, we might end up eliminating a team that is part of the true organization design if the value from retaining it in terms of the GMM loss is not sufficiently high given our sample.

By integrating the shape-constraints from Equation 15 and the penalty from Equation 16 into the usual GMM objective of Equation 14, we are now ready to characterize our proposed penalized and shape-constrained GMM estimator as follows:

$$\hat{\alpha} = \underset{\alpha \in \widetilde{\mathcal{A}}}{\arg\min}\ Q_n(\alpha) \equiv Q_n^0(\alpha) + \lambda_n \sum_{g \in \mathcal{G}^*} \sqrt{p_g} \left\| a_{g\cdot} \right\|_2 . \tag{17}$$

Given the empirical motivation in Section 1, our primary objective is to obtain consistent estimates of the production function rather than to conduct inference on the individual parameters. Consequently, we need not assume that the model presented in Section 3 fully captures the true underlying economic reality. Instead, we can view it as one of many potential representations of firm behavior. Any economic model that can be rationalized by our polynomial specification, subject to our parameter restrictions and sparsity, is acceptable within our framework.

4.3 Consistency of the Proposed Estimator and the Counterfactual Quantities

Adapting the arguments in Caner (2009) to group-LASSO, we are able to obtain consistency for our proposed estimator. We first state the assumptions which are standard in LASSO and GMM settings. Note that $\psi_i(\alpha) \equiv \rho\left(X_i, \alpha\right) * Z_i$ as in Equation 12.

Assumption 4.2.

For all $1 \leq i \leq n, n \geq 1$,

(a) $\psi_i(\alpha)$ is identically and independently distributed;

(b) $\left|\psi_i\left(\theta_1\right) - \psi_i\left(\theta_2\right)\right| \leq B_i \left|\theta_1 - \theta_2\right|$

with $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left[B_i^d\right] < \infty$, for some $d > 2$;

(c) $\sup_{\alpha \in \mathcal{A}} \mathbb{E}\left[|\psi_i(\alpha)|^d\right] < \infty$
for some $d > 2$.

Assumption 4.3.

Define $\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \psi_i(\alpha)\right] = m_{1n}(\alpha)$.

(a) Assume $m_{1n}(\alpha) \to m_1(\alpha)$ uniformly over $\mathcal{A}$, $m_{1n}(\alpha)$ is continuously differentiable in α, $m_1(\theta_0) = 0$, and $m_1(\alpha) \neq 0$ for $\alpha \neq \theta_0$; $m_1(\alpha)$ is continuous in α.

(b) Define the following $\mathcal{J} \times p$ matrix: $R_n(\alpha) = \partial m_{1n}(\alpha)/\partial\alpha'$. Assume that

$$R_n(\alpha) \xrightarrow{p} R(\alpha)$$

uniformly in a neighborhood N of θ_0. Here $R(\theta_0)$ is of full column rank, and $R(\alpha)$ is continuous in α.

Assumption 4.4. $W_n(\alpha)$ is a positive definite matrix that is continuous in $\alpha \in \mathcal{A}$, and $W_n(\alpha) \xrightarrow{p} W(\alpha)$ uniformly in α. The matrix $W(\alpha)$ is a symmetric nonrandom $\mathcal{J} \times \mathcal{J}$ matrix that is continuous in α and is positive definite for all $\alpha \in \mathcal{A}$.

We are now ready to state the consistency result for the proposed estimator.

Theorem 1 (Consistency of the Proposed Estimator). *Under Assumptions 1-3,*

1. *If $\lambda_n/n \to \lambda_0 \geq 0$, then uniformly in α*

$$\hat{\alpha}_n \xrightarrow{p} \arg\min_{\alpha \in \Gamma} Z(\alpha),$$

$$\text{where } Z(\alpha) = m_1(\alpha)' W(\alpha) m_1(\alpha) + \lambda_0 \sum_{g=1}^{G} \left\| \alpha^{(g)} \right\|_2 .$$

2. *If $\lambda_n = o(n)$,*

$$\hat{\alpha}_n \xrightarrow{p} \alpha_0.$$

Our counterfactual of interest is: What would the distribution of the quality of care across the universe of nursing homes in the U.S. have looked like in the year 2019 (the last year of our panel data) if minimum staffing requirements had been enforced then? Let M_{it} be the health outcome of interest, which as we argued before, is successful discharge rate for short term care quality or rate of ADL decline for long term care quality. We assume that M_{it} depends on the staffing vector L_{it} and firm attributes X_{it} through the following conditional distribution specification:

$$M_{it} \mid L_{it}, X_{it} \sim \text{Beta}(\alpha_{it}, \beta_{it}), \tag{18}$$

where α_{it} and β_{it} are the shape parameters of the Beta distribution, which are functions of the staffing vector L_{it} and firm-level attributes X_{it}. Specifically, we assume a log-linear specification for these parameters:

$$\alpha_{it} = \exp(Z_{it}'\pi_1), \quad \beta_{it} = \exp(Z_{it}'\pi_2), \tag{19}$$

where $Z_{it} = (L_{it}, X_{it})$ is the combined vector of covariates, and π_1 and π_2 are parameter vectors to be estimated. This non-linear link function ensures that the shape parameters α_{it} and β_{it} are strictly positive, which is required for the Beta distribution. This distribution is suitable for modeling rates, as it is defined on the unit interval and can capture different shapes of the underlying distribution. The probability density function of the Beta distribution for a given observation M_{it} is:

$$f(M_{it} \mid \alpha_{it}, \beta_{it}) = \frac{\Gamma(\alpha_{it} + \beta_{it})}{\Gamma(\alpha_{it})\Gamma(\beta_{it})} M_{it}^{\alpha_{it}-1}(1 - M_{it})^{\beta_{it}-1}, \tag{20}$$

where $\Gamma(\cdot)$ denotes the Gamma function. Given a sample of n firms observed over T time periods, the log-likelihood function is written as:

$$\begin{aligned} \ell(\pi_1, \pi_2) = \sum_{i=1}^{n} \sum_{t=1}^{T} \Big[&\log \Gamma(\alpha_{it} + \beta_{it}) - \log \Gamma(\alpha_{it}) - \log \Gamma(\beta_{it}) \\ &+ (\alpha_{it} - 1) \log M_{it} + (\beta_{it} - 1) \log(1 - M_{it}) \Big], \end{aligned} \tag{21}$$

where $\alpha_{it} = \exp(Z_{it}'\pi_1)$ and $\beta_{it} = \exp(Z_{it}'\pi_2)$. We can use MLE to estimate the parameter vectors π_1 and π_2 by maximizing the log-likelihood function as follows:

$$(\hat{\pi}_1, \hat{\pi}_2) = \arg\max_{\pi_1, \pi_2} \ell(\pi_1, \pi_2). \tag{22}$$

We have thus been able to consistently estimate the distribution of successful discharge rates M conditional on the staffing vector L and firm-time attributes X i.e $\hat{G}(M_{it} \mid L_{it}, X_{it}) \xrightarrow{p} G_0(M_{it} \mid L_{it}, X_{it})$ as $\hat{\pi} \xrightarrow{p} \pi_0$, where G is the Beta distribution, and $\pi \equiv (\pi_1', \pi_2')'$. The estimated shape parameters of the Beta Distribution are:

$$\hat{\alpha}_{it} = \exp(Z_{it}'\hat{\pi}_1), \quad \hat{\beta}_{it} = \exp(Z_{it}'\hat{\pi}_2). \tag{23}$$

Next, we obtain the post policy staffing mix of each firm. We suppress the t subscript in the rest of this subsection with the understanding that all variables are taken to be at their 2019 levels. We obtain the before-tax wage schedule w_i faced by the nursing home i from the US-Bureau of Labor Statistics. It varies only at the state level but does show considerable variation across states due to sticky local labor market conditions. Firm i solves the following cost minimization problem:

$$L_i^* = \underset{L \in \mathbb{R}_+^d}{\arg\min} \;\; w_i'L \quad \text{subject to } L \geq \bar{L}_i, \text{ and } h(L;\theta) + \beta_k k_i + \omega_i = y_i - \varepsilon_i$$

where $\bar{L}_i$ captures the minimum staffing requirements for the labor vector. Note that $\omega_i + \varepsilon_i$ is simply equal to $y_i - h(L_i;\theta) - \beta_k k_i$. Substituting that, we get:

$$L_i^* = \underset{L \in \mathbb{R}_+^d}{\arg\min} \;\; w_s'L \quad \text{subject to } L \geq \bar{L}, \text{ and } h\left(L;\theta\right) = h\left(L_i;\theta\right) \tag{24}$$

We compute the above at $\hat{\theta}$ obtained from our proposed estimator. Let $\{(L_i^*, X_i)\}_{i=1}^n$ represent the post-policy staffing vector $L_i^* \in \mathbb{R}_+^d$ and firm attributes $X_i \in \mathbb{R}^p$. The empirical CDF estimator for the joint distribution of L_i^*, X_i can then be written as:

$$\hat{F}_{L^*,X}(L, X) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(L_i^* \leq L, X_i \leq X), \tag{25}$$

where $\mathbb{1}(\cdot)$ is the indicator function, which equals 1 if $L_i^* \leq L$ and $X_i \leq X$ hold component-wise, and 0 otherwise. It places equal probability $\frac{1}{n}$ on each observed pair (L_i^*, X_i). The estimator for the post-policy marginal distribution of M_{it} is:

$$\hat{F}_{M^*}(m) = \int \hat{\mathbb{E}}\left[\mathbb{1}(M_i \leq m) \mid L_i^*, X_i\right] \hat{F}_{L^*,X}\left(L^*, X\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_{M|L_i^*,X_i}(m), \tag{26}$$

where $\hat{F}_{M|L_i^*,X_i}(m) = \mathrm{BetaCDF}\left(m; \hat{\alpha}\left(L_i^*, X_i\right), \hat{\beta}\left(L_i^*, X_i\right)\right)$. If, instead, we were interested in estimating the post-policy marginal density, then the estimator is:

$$\hat{f}_{M^*}(m) = \int \hat{f}_{M|L,X}\left(m \mid L^*, X\right) d\hat{F}_{L^*,X}\left(L^*, X\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{M|L_i^*,X_i}(m), \tag{27}$$

where $\hat{f}_{M|L_i^*,X_i}(m) = \frac{\Gamma\left(\hat{\alpha}(L_i^*,X_i)+\hat{\beta}(L_i^*,X_i)\right)}{\Gamma(\hat{\alpha}(L_i^*,X_i))\Gamma\left(\hat{\beta}(L_i^*,X_i)\right)} m^{\hat{\alpha}(L_i^*,X_i)-1}(1-m)^{\hat{\beta}(L_i^*,X_i)-1}$, and $\hat{\alpha}(\cdot)$ and $\hat{\beta}(\cdot)$ are as in Equation 23. This is what we present in the empirical section.

5 MONTE-CARLO SIMULATIONS

Consider the case of two worker types: clinical nursing staff and administrative staff. The number of hours employed of each in firm i at time t is given by L_{1it} and L_{2it} respectively. The fixed capital assets are denoted by K_{it}. We consider a sample of $n = 1000$ facilities for $T = 3$ years, yielding a balanced panel with $3{,}000$ observations. With two worker types, we have 3 different teams under consideration: the nurse working alone, the admin working alone, and both working together. Then, the labor index $H(L;\theta_0)$ with all three of these teams represented is given by:

$$H(L;\theta_0) = v_1(a_1 L_1)^{\gamma_1} + v_2(a_2 L_2)^{\gamma_2} + v_3(a_{31}L_1)^{\gamma_{31}}(a_{32}L_2)^{\gamma_{32}}, \tag{28}$$

where $\theta_0 \equiv (a_1, a_2, a_{31}, a_{32}, \gamma_1, \gamma_2, \gamma_{31}, \gamma_{32})'$ is the true parameter vector that characterizes the labor index. Suppose both of the workers work alone and the team in which they work together is not deployed. So, $a_{31} = a_{32} = 0$ in Equation 28. Since the time budget constraint has to be satisfied, this implies that $a_1 = a_2 = 1$. To understand the data generating process, let us start with the first time period $t = 1$. For this year, for each of the variables L_1, L_2 and K, we take out n i.i.d. draws from their respective empirical distributions for the year 2019. For each nursing home, we draw the idiosyncratic shock ε_{it} from a mean zero normal distribution with variance σ_ε and assume that productivity follows an AR(1) process given by:

$$\omega_{it+1} = \delta_1\omega_{it} + \xi_{it+1}, \tag{29}$$

where ξ_{it} is an innovation shock which we assume has a mean zero normal distribution with variance σ_ξ. Then, for this $t = 1$, we initialize productivity of each nursing home to be drawn i.i.d. from its stationery distribution, which, from the normality of ξ and from Equation 29, is also a mean zero normal distribution with variance $\sigma_\xi/(1-\delta_1)$. Then, the output (in logs) of nursing home i at time $t=1$ is:

$$y_{it} \equiv \log F(X_{it};\theta_0,\beta_0) + \varepsilon_{it} = \delta_0 + \log H(L_{1it}, L_{2it};\theta_0) + \beta_k k_{it} + \omega_{it} + \varepsilon_{it}, \quad (30)$$

where $X_{it} \equiv (L'_{it}, K_{it}, \omega_{it})'$, and $\beta_0 = (\delta_0, \beta_k)'$. For the next time period $t=2$, we assume that capital evolves according to the following law of motion:

$$K_{it+1} = (1-\nu)K_{it} + I_{it}, \quad (31)$$

where ν is the depreciation rate of the capital stock and I_{it} is the fresh investment made by the firm in time t which is realized after a gestation lag of 1 time period. We assume that this investment has the following functional form representation:

$$I_{it} = (\nu + \kappa \cdot \omega_{it})K_{it}. \quad (32)$$

Next, for each firm i, we draw their innovation shock from a mean zero normal distribution with variance σ_ξ^2. Recall that we assumed productivity ω_{it+1} to evolve according the an AR(1) process of the form given in Equation 29. Labor is flexible and chosen to minimize short run costs subject to staying on the same isoquant:

$$\begin{aligned} L_{it+1} = \arg\min_{L\in\mathbb{R}_+^2} w'_{it+1}L \quad \text{subject to} \\ F\left((L', K_{it+1}, \omega_{it+1})'; \theta_0, \beta_0\right) = F\left((L'_{it}, K_{it}, \omega_{it})'; \theta_0, \beta_0\right), \end{aligned} \quad (33)$$

where the wage schedule evolves with lognormal correlated shocks as follows:

$$\log w_{it+1} = \alpha + \rho \log w_{it} + \Sigma^{1/2}\epsilon_{it+1},$$

where α is a vector of intercepts representing the overall wage growth, ρ is a diagonal matrix representing the persistence of wages for each worker type, $\Sigma^{1/2}$ is the Cholesky decomposition of the covariance matrix Σ, capturing the correlations between the wages, and $\epsilon_{it+1} \sim \mathcal{N}(0, \mathrm{I})$ is a vector of independent standard normal shocks. Note that the labor choice depends on the current productivity ω_{it+1} which means it will be endogenous. Also, we need some exogenous variation in labor for it to be identified. That is happening through the shifting of the wage schedule. The beginning wage for $t = 1$ we use to kick-start the evolution is the wage rate prevailing in 2019 obtained from the US Bureau of Labor Statistics.

As before, we draw idiosyncratic shocks $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and obtain the output as in Equation 30. We can get the data for all the later years recursively from $t = 2$. The list of all the parameters with their true values is given in Table 3 below.

Table 3: Chosen Parameter Values for the Monte-Carlo Experiment

Parameter	**True Value**	**Description**
a_{11}	1.0	allocation of nurse in team (nurse)
γ_{11}	0.7	elasticity of nurse in team (nurse)
v_1	2.0	productivity of team (nurse)
a_{22}	1.0	allocation of admin in team (admin)
γ_{22}	0.9	elasticity of admin in team (admin)
v_2	1.0	productivity of team (admin)
a_{31}	0.0	allocation of nurse in team (nurse,admin)
a_{32}	0.0	allocation of admin in team (nurse,admin)
γ_{31}	0.4	elasticity of nurse in team (nurse,admin)
γ_{32}	0.6	elasticity of admin in team (nurse,admin)
v_3	0.5	productivity of team (nurse,admin)
β_k	0.3	output elasticity of capital
δ_0	0.0	constant (firm and time invariant)
δ_1	0.7	autoregressive coefficient
σ_ε	0.01	variance of idiosyncratic shock
σ_ξ	0.1	variance of innovation shock
ν	0.1	depreciation rate of capital stock
κ	0.3	investment coefficient
α_1	0.02	long run growth rate of nurse wages
α_2	0.015	long run growth rate of admin wages
ρ_1	0.9	persistence of nurse wages
ρ_2	0.7	persistence of admin wages
Σ_{11}	0.01	volatility in nurse wages
Σ_{22}	0.005	volatility in admin wages
Σ_{12}	0.003	strength of correlation in wage shocks

We are interesting in estimating not just the production function F, but certain functionals of it, such as the direct partial elasticity of substitution. For any pair of inputs (j_1, j_2) evaluated at the input vector x, it has the following expression:

$$\sigma_{j_1 j_2}(x) = \frac{F_1' F_2' \left(x_1 F_1' + x_2 F_2'\right)}{x_1 x_2 \left[2F_1' F_2' F_{12}'' - \left(F_2'\right)^2 F_{11}'' - \left(F_1'\right)^2 F_{22}''\right]}, \tag{34}$$

where F_1' and F_1'' are the first and second order derivatives respectively with respect to j_1, and F_{12}'' is the cross partial derivative with respect to j_1 and j_2. It is analogous for F_2' and F_2''. All of these are evaluated at the input vector x. Thus, this formula for elasticity is measured keeping the other inputs fixed at some given level. Note that the Hicks neutrality assumption baked into Equation 30 implies that we do not need to know what the unobserved productivity is to compute Equation 34 as it would simply drop out. In multi-factor production functions, there are other ways of measuring this, such as the Allen and shadow partial elasticity of substitutions. Literature on this goes back to the 1960s with work by McFadden (1963), later consolidated in Silberberg (1978) and Sydsæter, Strøm, and Berck (2005).

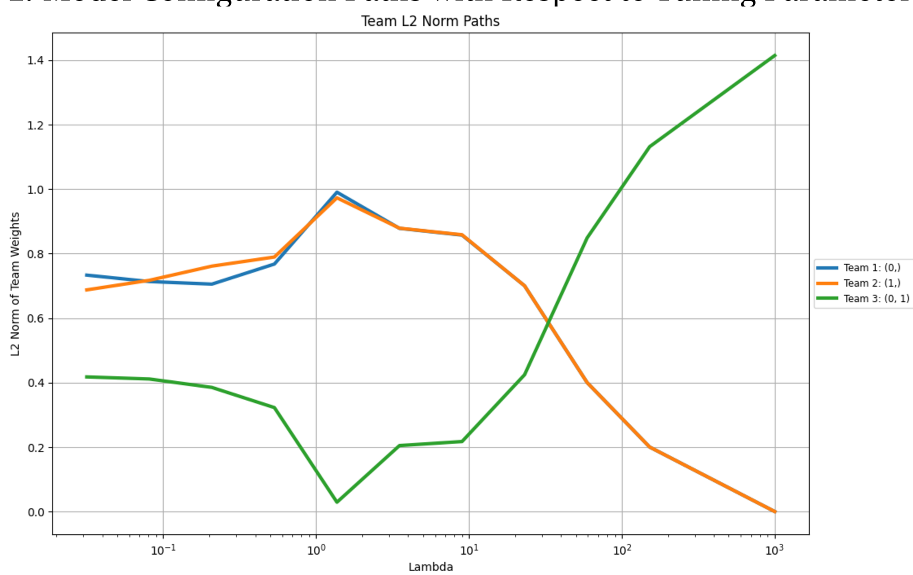
We will consider four different models for estimation. The first one is when we simply know the true model and estimate the parameters using standard GMM. Second is when we consolidate the two worker types into a single index $L_{it} = L_{1it} + L_{2it}$ and run the GMM estimator with a Cobb-Douglas specification. between L_{it} and K_{it}. Third, we use the LASSO criterion with a Bernstein polynomial specification. Finally, we estimate using our proposed estimator which we dub as PSC-GMM (Penalized and Shape Constrained GMM). We conduct $S = 1000$ simulations. The results are summarized in Table 4 below. We also plot the ℓ_2-norm of the team allocation coefficients as a function of the tuning parameter in Figure 2, referred to

Table 4: RMSE for the Production Function and Elasticities Across Models

Model	**RMSE**			
	H	$\boldsymbol{\sigma_{L_1 L_2}}$	$\boldsymbol{\sigma_{L_1 K}}$	$\boldsymbol{\sigma_{L_2 K}}$
True Specification Known				
Aggregated Index + Cobb-Douglas				
Bernstein Polynomial + LASSO				
Proposed Estimator (PSC-GMM)				

as model configuration paths. These paths are considerably stable across different tuning parameter values. Note the interaction between the time budget constraint and the tuning parameter. When the tuning parameter is set too high, the estimator tends to select the team with both the workers together, even though this is not the correct specification. This occurs because the reduction in the GMM loss is outweighed by the increase in costs associated with selecting more teams.

Figure 2: Model Configuration Paths with Respect to Tuning Parameter Choice



6 IMPACT OF MINIMUM STAFFING MANDATES ON LABOR AND WELFARE

6.1 Estimation Results

First, we consider the case in which we do not treat for any endogeneity and simply run an OLS regression while retaining the proposed specification, shape constraints and penalty. We first determine the value of the tuning parameter λ_{GMM} using K-fold cross validation for the GMM loss. Then, we find a corresponding value of the tuning parameter in the OLS setting which selects the same number of teams as we do under GMM so that the results are comparable. This turns out to be $\lambda_{OLS} = 0.1$. Out of the 31 teams, only 8 were selected, with our regularization term shrinking the allocation vectors corresponding to the remaining 23 teams to zero. The parameter estimates for the surviving teams are given in Table 5 below:

Team	**Allocation (a)**	**Elasticity (γ)**	**Match Surplus (v)**
admin, therapist	0.55, 0.51	0.51, 0.49	4.15
specialist, therapist	0.64, 0.49	0.50, 0.50	2.75
admin	0.44	1.00	2.26
nurse	0.51	1.00	2.05
specialist	0.36	1.00	2.04
nurse, wellness	0.49, 0.79	0.54, 0.46	1.95
wellness	0.20	1.00	1.65
admin, wellness	0.01, 0.01	0.58, 0.42	1.03

Table 5: Parameter Estimates of the Surviving Teams Under OLS

In the table, the selected teams are ranked according to the match surplus they are associated with. The results from the OLS method are quite puzzling. Most of the teams are very small. The nurse, in particular, only forms a team with the wellness staff. The time-team elasticities are always hitting unity, suggesting that this constraint binds for all teams as all of them are exhibiting constant returns to scale. It is also surprising for the nurse and specialists to not be in as many high value teams. All of this suggests that the problem of endogeneity is likely quite severe. The estimates of the auxiliary parameters: the constant and the output elasticity of capital, can be found in Table 7 below. The output elasticity of capital is abnormally low. This is in line with other studies that suggest that in the presence of endogeneity, we often end up underestimating this parameter when using OLS.

Next we run the proposed penalized and shape constrained GMM procedure as laid out in Section 4. The estimates we obtain for the surviving teams are in Table 6 and those for the auxiliary parameters (output elasticity of capital, constant, and the auto-regressive coefficient) are recorded in Table 7. Some more details on both the OLS and GMM results can be found in **??**.

The results are much more sensible now. The total number of selected teams is 8, which is one-quarter of the set of all possible team configurations. The average team size is 2, with a standard deviation of 0.87, indicating a moderate variability in the distribution of team sizes. The administrative staff is the most dispersed, participating in 5 teams. On the other hand, the specialist is the most focused, be-

Table 6: Parameter Estimates of the Surviving Teams Under GMM

Team	**Allocation (a)**	**Elasticity (γ)**	**Match Surplus (v)**
nurse, wellness, specialist	0.15, 0.09, 0.76	0.34, 0.27, 0.10	1.87
wellness	0.81	0.99	1.76
admin, therapist	0.07, 0.22	0.39, 0.45	1.72
admin, nurse, wellness	0.20, 0.06, 0.10	0.59, 0.10, 0.20	1.60
admin	0.57	1.00	1.33
admin, specialist, therapist	0.06, 0.24, 0.16	0.38, 0.16, 0.21	1.29
admin, nurse	0.10, 0.79	0.10, 0.80	0.85
therapist	0.62	0.89	0.25

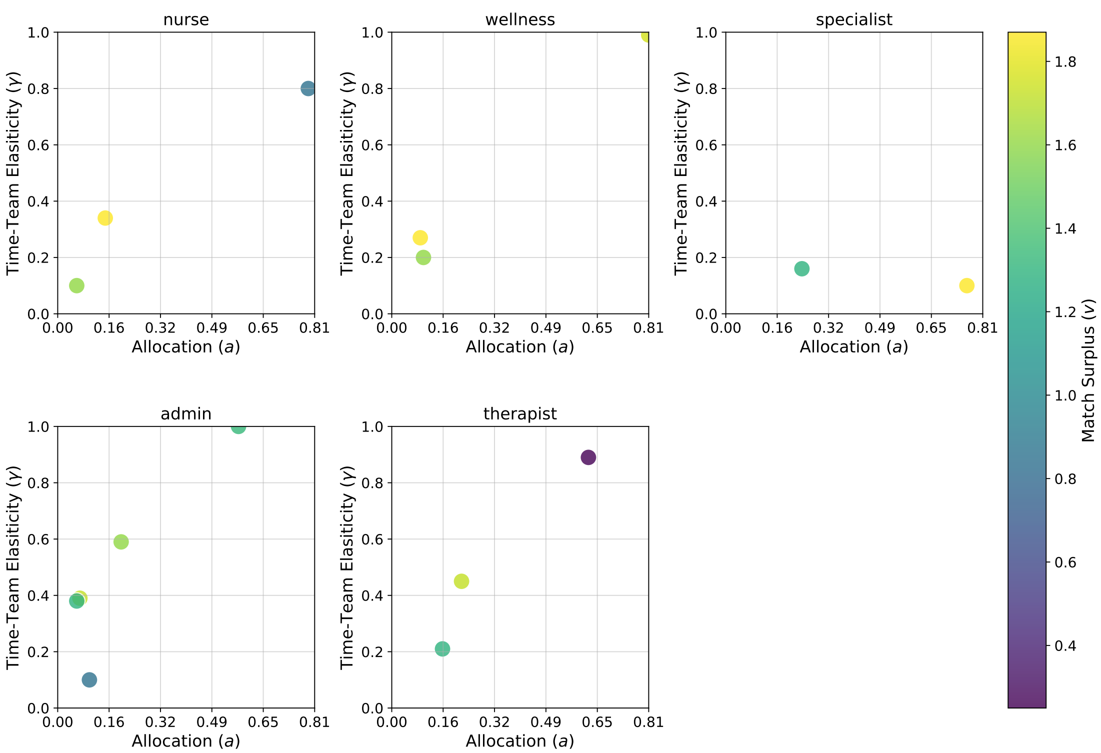
Table 7: Estimates of Auxiliary Parameters Under OLS and GMM

Parameter	**Estimated Value (OLS)**	**Estimated Value (GMM)**
β_k	0.14	0.35
δ_0	2.71	1.26
δ_1	-	0.67

ing in just 2 teams. However, those are both large teams and around 80% of the specialist's hours are devoted to the team with the high-match surplus. This can be driven by (1) the team having a high skill level and performing complex tasks tasks, or (2) the density of tasks allocated to this team is very high i.e. they are performing many tasks. There is an average overlap of 0.71 worker types per pair of teams. Teams having common members implies a certain level of redundancy or shared responsibilities, which could be a strategy for ensuring coverage and flexibility within the organization. Although the goal of this paper is not inference on the individual parameters or how they might identify certain primitives of the economic model, it is encouraging that the estimates are economically interpretable and align with our qualitative understanding of the organization design.

Figure 3 visualizes the estimated parameters for each worker type. Each point represents a team, with the color indicating the team's match surplus, and the position reflecting the worker's estimated allocation and team-time elasticity parameters. The fact that the points are quite dispersed and not clustered suggests that the role and impact of each worker type are highly team-dependent. The monotonic relationship between allocation and team-time elasticity suggests that workers who

Figure 3: Parameter Plots by Worker Types



allocate more hours to a team also contribute more to its overall effective time. There is no clear link between allocation and surplus— nurses, for example, may allocate significant time to teams with lower surplus, as their tasks, such as patient monitoring, do not always drive revenue. Specialists, however, are more often involved in higher-surplus teams due to the targeted nature of their contributions.

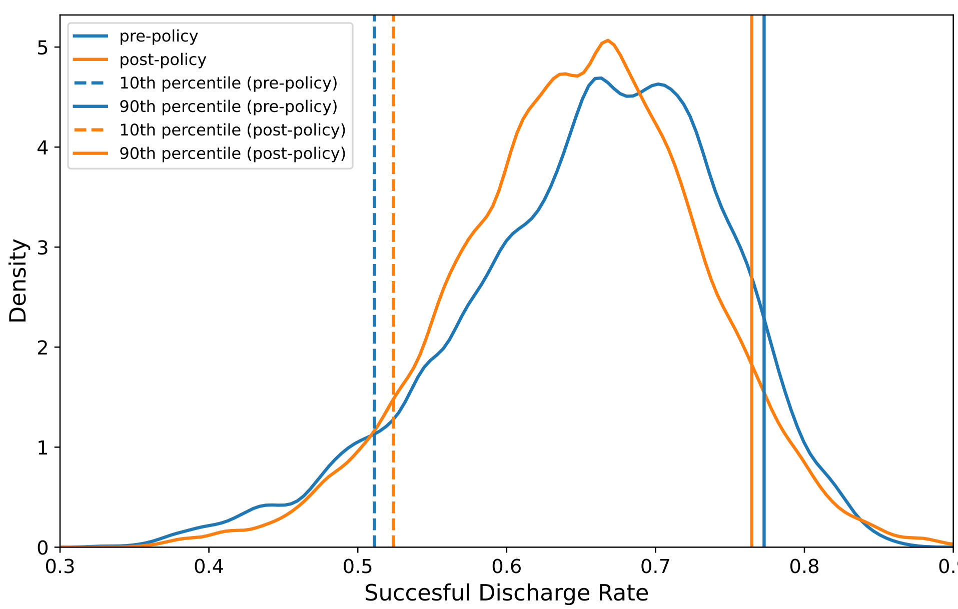
6.2 Counterfactual Results

On April 22, 2024, the Centers for Medicare & Medicaid Services (CMS) issued new minimum staffing standards, which will have a staggered rollout over the next five years all across the United States. The standards stipulate a total nurse staffing requirement of 3.48 hours per resident day (HPRD), a threshold that poses a binding constraint for 80% of nursing facilities nationwide. It further specifies that out of the total 3.48 HPRD, 0.55 HPRD must be provided by registered nurses (RNs) and

2.45 HPRD by nurse aides (NAs). The remaining 0.48 HPRD can be fulfilled through any combination of RNs, licensed practical nurses (LPNs), and NAs, offering some flexibility for facilities to adapt their clinical nurse staffing configurations according to their specific needs and resources.

Since we lump all nurses into the clinical nursing worker type for our empirical exercise, we will primarily be imposing the overall constraint of 3.4 HPRD on that group. In our data, we see that 60% of nursing hours come from NAs, while RNs and LPNs contribute 20% each. Therefore, when we implement the minimum staffing threshold, we implicitly assume that the hours for each worker type in the targeted group increase proportionally, maintaining the existing staff composition. In other words, we implicitly require that NAs meet 2.1 HPRD, while RNs and LPNs each meet 0.65 HPRD. This is not too far off from the actual policy requirements.

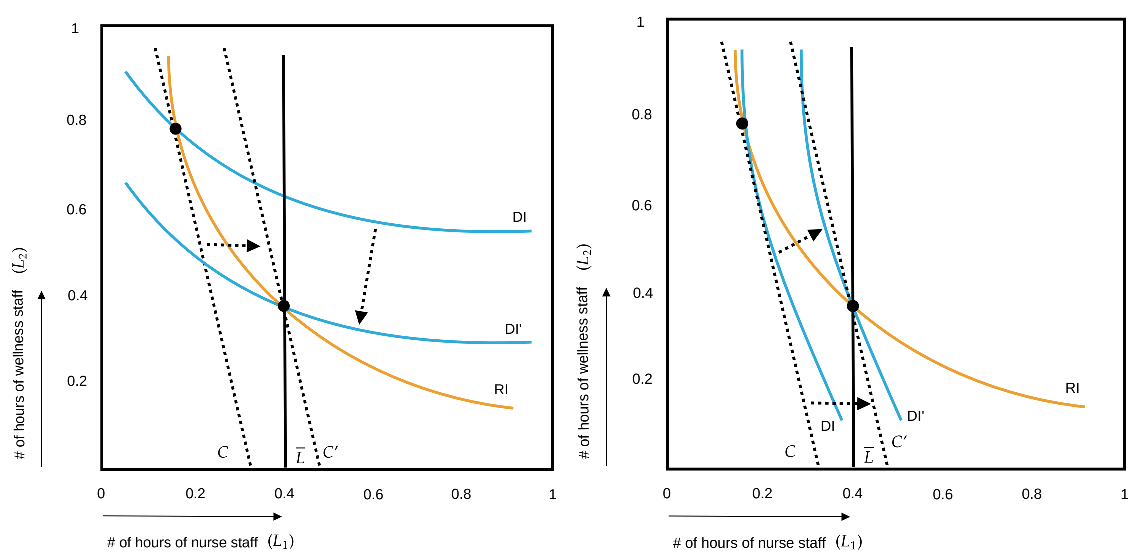
Figure 4: Pre and Policy Policy Density of Successful Discharge Rate



We abstract away from issues of non-compliance, since CMS hasn't yet formally laid out the repercussions if a facility fails to meet these standards in a stipulated timeframe. If the failure stems from exogenous conditions like a tight local labor market, then the firms are granted temporary exceptions. There are also many infractions that are dealt with on a case-by-case basis. We assume for now that firms are able to adjust their staffing mix at the prevailing market wage rate. Following the steps leading upto Equation 27, we plot the pre and post policy marginal density of the successful discharge rate in the population of nursing homes in Figure 4. As discussed before, it is appropriate to measure short-term quality of care.

We see an improvement in this rate by about 2 percentage points among the lower decile firms as well as drop in the rate among the upper decile by about 1 percentage point. These are economically significant effects since in absolute numbers they suggest difference in discharges of thousands of patients per year. The mean rate fell while the disparity between the facilities also reduced. These trade-offs render common decision criteria— such as first and second order stochastic dominance, inadequate to be able to conclude whether the policy is welfare enhancing.

Figure 5: Welfare Depends on the Curvature of Production and Health Isoquants



(a) Policy is Welfare Enhancing

(b) Policy is Welfare Reducing

The intuition as to why the policy improves health outcomes for some facilities while reducing for the others comes from the first principles of producer theory as illustrated in Figure 5 above. RI is is the revenue isoquant and DI is the successful discharge rate isoquant. On the axes we have 2 worker types: nurse and wellness staff. If the health isoquant is flatter than the revenue isoquant, then movement along the right on RI will cause us to be on a lower DI', as seen in the left panel. On the other hand, if DI is steeper than RI in the adjustment region, then a movement along with right on RI will put us on a higher DI', as seen in the right panel. In our counterfactual, we have 5 dimensional isoquants which can have quite complicated curvatures in different regions. So it is easy to see why some nursing homes might exhibit movement as in the left panel and others as in the right.

7 CONCLUSION AND NEXT STEPS

In this paper, we develop a disaggregated, flexible, and parsimonious model of team based production and propose a novel penalized and shape-constrained GMM estimator for it in the context of the nursing home industry. Using theories from personnel economics and organizational design, we specify our model and the restrictions and sparsities on its parameters. The organisation design is not known to the researcher and is informed from the data subject to (1) the time budget constraint, which ensures all workers' hours are accounted for, and (2) costly team formation, with the cost increasing with both the number of teams and their size.

Our estimator selects 8 out of the 31 possible team configurations of five different worker types: administrative staff, clinical nursing staff, specialist, wellness staff, and therapist. Both very large and very small teams are less likely to occur. Specialists are the most concentrated— working in only a few but high value large teams. On the other hand, the administrative staff is the most diffuse and agnostic to team size. In our counterfactual, we predict the distribution of the short term health outcomes as measured by the successful discharge rate in the universe of nursing homes in the year 2019 under the minimum staffing requirements. This analysis is timely, as CMS plans to implement these staffing mandates nationwide from 2024 through 2029. We see an improvement in this rate by about 2 percentage points among the lower decile firms as well as drop in the rate among the upper decile by about 1 percentage point. These numbers are economically significant as they suggest a difference to the tune of thousands of patients per year. The mean rate fell while the dispersion also reduced. These trade-offs render common decision criteria used by policymakers— such as first and second order stochastic dominance, inadequate to be able to conclude whether the policy is welfare enhancing. A rigorous welfare analysis however, would require us to accommodate general equilibrium effects along with a more realistic functioning of the input and output markets. For example, this policy might affect the demand for nursing homes by alleviating some information frictions that hamper the patients when selecting into a facility of their choice. This is a promising avenue for future research.

Beyond the nursing home setting, this methodology can be useful in many areas of interest to applied researchers. For example, in skill formation models, it provides a robust way to incorporate domain-specific institutional knowledge about how different inputs (monetary, temporal, home, and school-based) interact, while remaining agnostic about the exact grouping structure, letting it be data-driven.

An important extension we are currently exploring embeds our framework in a semi-nonparametric setting, with three key benefits. First, it allows us to use nonparametric estimation techniques for auxiliary functions, expanding the set of identification strategies available to the researcher. Second, it enables us to leverage all the information in the conditional moment restrictions, improving efficiency and addressing identification concerns. Third, it accommodates a high-dimensional regime in which the number of inputs grow with the sample size. Additionally, we are interested in the estimation and inference of model primitives, such as the skill embeddings of worker types and the tasks they perform. This could help identify skill mismatches with important policy implications on the design of on-the-job training programs and the potential for task automation.

REFERENCES

Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* 59 (2): 391–425.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." Working Paper 16082, National Bureau of Economic Research.

Ackerberg, Daniel A. 2023. "Timing Assumptions and Efficiency: Empirical Evidence in a Production Function Context." *The Journal of Industrial Economics* 71 (3): 644–674.

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

Ackerberg, Daniel, Jinyong Hahn, and Qingsong Pan. 2022. "Nonparametric Identification Using Timing and Information Set Assumptions with an Application to Non-Hicks Neutral Productivity Shocks." *Working Paper*.

Almagro, Milena, and Elena Manresa. 2021. "Data-Driven Nests in Discrete Choice Models."

Blundell, Richard, and Stephen Bond. 2000. "GMM Estimation with persistent panel data: an application to production functions." *Econometric Reviews* 19 (3): 321–340.

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen. 2007. "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves." *Econometrica* 75 (6): 1613–1669.

Bonhomme, Stephane. 2021. "Teams: Heterogeneity, Sorting, and Complementarity."

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica* 83 (3): 1147–1184.

Burton, Richard M., Børge Obel, and Dorthe Døjbak Håkonsson. 2021. *Organizational Design: A Step-by-Step Approach*. 4th ed.: Cambridge University Press.

Caner, Mehmet. 2009. "Lasso-Type GMM Estimator." *Econometric Theory* 25 (1): 270–290.

Caner, Mehmet, Xu Han, and Yoonseok Lee. 2018. "Adaptive Elastic Net GMM Estimation With Many Invalid Moment Conditions: Simultaneous Model and Moment Selection." *Journal of Business & Economic Statistics* 36 (1): 24–46.

Caplin, Andrew, David J Deming, Søren Leth-Petersen, and Ben Weidmann. 2023. "Allocative Skill." Working Paper 31674, National Bureau of Economic Research.

Chernozhukov, Victor, Whitney K. Newey, and Andres Santos. 2023. "Constrained Conditional Moment Restriction Models." *Econometrica* 91 (2): 709–736.

Chetverikov, Denis, and Daniel Wilhelm. 2017. "Nonparametric Instrumental Variable Estimation Under Monotonicity." *Econometrica* 85 (4): 1303–1320.

Chidambaram, Priya, and Alice Burns. 2022. "10 Things About Long-Term Services and Supports (LTSS)." Accessed: 2024-06-08.

Ching, Andrew T., Fumiko Hayashi, and Hui Wang. 2015. "QUANTIFYING THE IMPACTS OF LIMITED SUPPLY: THE CASE OF NURSING HOMES." *International Economic Review* 56 (4): 1291–1322.

Demirer, Mert. 2020. "Essays on Productivity, Misallocation, and Firm Dynamics." Job Market Paper, University of Pennsylvania, Department of Economics.

Dessein, Wouter, Desmond (Ho-Fu) Lo, and Chieko Minami. 2022. "Coordination and Organization Design: Theory and Micro-Evidence." *American Economic Journal: Microeconomics* 14 (4): 804–43.

Gandhi, Ashvin. 2023. "Picking Your Patients: Selective Admissions in the Nursing Home Industry." (3613950).

Gillen, Benjamin J, Sergio Montero, Hyungsik Roger Moon, and Matthew Shum. 2019. "BLP-2LASSO for aggregate discrete choice models with rich covariates." *The Econometrics Journal* 22 (3): 262–281.

Grabowski, David C, Zhanlian Feng, Richard Hirth, Momotazur Rahman, and Vincent Mor. 2013. "Effect of nursing home ownership on the quality of post-acute care: an instrumental variables approach." *Journal of Health Economics* 32 (1): 12–21. Epub 2012 Sep 14.

Gupta, Atul, Sabrina T Howell, Constantine Yannelis, and Abhinav Gupta. 2023. "Owner Incentives and Performance in Healthcare: Private Equity Investment in Nursing

Homes." *The Review of Financial Studies* 37 (4): 1029–1077.

Haag, Berthold R., Stefan Hoderlein, and Krishna Pendakur. 2009. "Testing and imposing Slutsky symmetry in nonparametric demand systems." *Journal of Econometrics* 153 (1): 33–50.

Hackmann, Martin B. 2019. "Incentivizing Better Quality of Care: The Role of Medicaid and Competition in the Nursing Home Industry." *American Economic Review* 109 (5): 1684–1716.

Kasahara, Hiroyuki, Paul Schrimpf, and Michio Suzuki. 2023. "Identification and Estimation of Production Function with Unobserved Heterogeneity."

Lazear, Edward P., and Michael Gibbs. 2014.

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

Liao, Zhipeng. 2013. "Adaptive Gmm Shrinkage Estimation with Consistent Moment Selection." *Econometric Theory* 29 (5): 857–904.

Lin, Haizhen. 2015. "Quality Choice and Market Structure: A dynamic analysis of nursing home oligopolies." *International Economic Review* 56 (4): 1261–1290.

Lindenlaub, Ilse. 2017. "Sorting Multidimensional Types: Theory and Application." *The Review of Economic Studies* 84 (2): 718–789.

Matzkin, Rosa L. 2013. "Nonparametric Identification in Structural Economic Models." *Annual Review of Economics* 5 (1): 457–486.

McFadden, Daniel. 1963. "Constant Elasticity of Substitution Production Functions." *The Review of Economic Studies* 30 (2): 73–83.

Menzel, Konrad. 2022. "Structural Sieves."

Ocampi, Sergio. 2022. "A Task Based Theory of Occupations."

Olenski, Andrew. 2023. "Reallocation and the (In)efficiency of Exit in the U.S. Nursing Home Industry."

Olley, G Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64 (6): 1263–97.

Silberberg, E. 1978. *The Structure of Economics: A Mathematical Analysis.*: McGraw-Hill.

Su, Liangjun, Zhentao Shi, and Peter C. B. Phillips. 2016. "Identifying Latent Structures in Panel Data." *Econometrica* 84 (6): 2215–2264.

Sydsæter, Knut, Arne Strøm, and Peter Berck. 2005. *Economists' Mathematical Manual.* 4th ed. Berlin, Heidelberg: Springer.

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

Wang, Ao. 2023. "Sieve BLP: A semi-nonparametric model of demand for differentiated products." *Journal of Econometrics* 235 (2): 325–351.

Yuan, Ming, and Yi Lin. 2006. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1): 49–67.

A ALTERNATIVE IDENTIFICATION STRATEGIES

We first introduce another set of inputs called materials which enter the gross production function. Consider again the value-added production (in logs) in **??**:

$$y_{it} = h_0 \left(L_{it}, K_{it}\right) + \omega_{it} + \varepsilon_{it} \tag{35}$$

Assumption A.1 (Timing of Input Selection)**.**

1. Materials M_{it} are perfectly flexible inputs and chosen at time t.
2. Capital K_{it} is a quasi-fixed input and chosen at time $t-1$.
3. Partially flexible labor L_{it} chosen between capital & materials at $t-b$, $b \in (0,1)$.

Materials maximize flow profits/minimize short run production costs. Capital with dynamic implications (LOM: $K_{it} = \delta K_{it-1} + I_{it-1}$). **??** requires capital and labor to be chosen before current period productivity is drawn but allows their choice to have dynamic considerations $\big($like capital following a law of motion of the form $K_{it} = \gamma K_{it-1} + I_{it-1}\big)$ while abstracting away from the need to model it.

Assumption A.2 (Properties of the Unobserved Productivity)**.**

1. Productivity ω_{it} is Markovian, so $\omega_{it} = h_1\left(\omega_{it-1}\right) + \xi_{it}$, where $\mathbb{E}\left[\xi_{it} \mid I_{it-1}\right] = 0$.
2. Intermediate input demand $m_{it} \equiv \psi_t(\ell_{it}, k_{it}, \cdot)$ is strictly monotonic in ω_{it}.

Let $V_t \equiv (X_t, Z_t)$ be the vector of all random variables, with $X_t \equiv \left(W_t, Z_t^*\right)$. Here $W_t = (y_{t+1}, \ell_{t+1})$ is endogenous while $Z_t^* = (y_t, \ell_t, k_t, k_{t+1})$ is exogenous and is a subset of all the instruments $Z_t \equiv I_t$. We get these $2T-1$ conditional moment restrictions:

$$\mathbb{E}\left[\rho\left(X, \alpha_0\right) \mid Z\right] \equiv \{\mathbb{E}\left[\rho_t\left(X_t, \alpha_0\right) \mid I_t\right]\}_{t=1}^{\mathcal{J}} = 0 \tag{36}$$

where $\rho_t(Z,\alpha) = (\varepsilon_t, \xi_{t+1})'$ so $d_{\rho_t} = 2$ for $t = 1, \ldots, T-1$ and $\rho_T(Z,\alpha) = \varepsilon_t$ so $d_{\rho_T} = 1$. Here, $\alpha \equiv (h_0, h_1, h_2) \in \prod_{i=1}^{3} \mathcal{H}_i \equiv \mathcal{A}$, where h_0 is the value added production function, h_1 is the markovian function describing the law of motion of productivity, and h_2 is a nuisance function. The residual functions ε_t and ξ_{t+1} are:

$$\varepsilon_t = y_t - h_2\left(\ell_t, k_t, m_t\right)$$
$$\xi_{t+1} = h_2\left(\ell_{t+1}, k_{t+1}, m_{t+1}\right) - h_0(\ell_{t+1}, k_{t+1}, \omega_{t+1}) - h_1\left(h_2\left(\ell_t, k_t, m_t\right) - h_0(\ell_t, k_t, \omega_t)\right)$$

B SIEVE INTERPRETATION

Suppose the labor index H emerges from the performance of M teams. We also assume that the output of team m is characterized by the CES technology ϕ_m as:

$$\phi_m\left(L; \theta_m^{(1)}\right) = \left[\sum_{j=1}^{d} \gamma_{mj} \alpha_{mj}^{\rho_m} L_j^{\rho_m}\right]^{\tau_m/\rho_m}, \text{ for } m = 1, \ldots, M,$$

where $\theta_m^{(1)} \equiv \left(\gamma_{m\bullet}, \alpha_{m\bullet}, \rho_m, \tau_m\right)$. Here, $\sigma_m = 1/(1-\rho_m)$ captures the substitutability between inputs in team m, γ_{mj} and α_{mj} are the relative productivity and the fraction of hours of input j in team m respectively. τ_m is the returns to scale in team m. Let $\phi\left(L; \theta^{(1)}\right) \equiv \left(\phi_1\left(L; \theta_1^{(1)}\right), \ldots, \phi_M\left(L; \theta_M^{(1)}\right)\right)$. These team outputs are combined to yield the labor index H using the CES technology ψ as follows:

$$H\left(X; \theta\right) \equiv \psi\left(\phi\left(L; \theta^{(1)}\right); \theta^{(2)}\right) = \left[\sum_{m=1}^{M} \gamma_m \left(\phi_m\left(L; \theta_m^{(1)}\right)\right)^{\rho}\right]^{\tau/\rho},$$

where $\theta^{(2)} \equiv \left(\gamma_\bullet, \rho, \tau\right)$ and $\theta \equiv \left(\theta^{(1)}, \theta^{(2)}\right)$. As before, $\sigma = 1/(1-\rho)$ captures the substitutability between teams, γ_m is the relatively productivity of team m, and τ is the returns to scale over the teams. By allowing M to be a tuning parameter than increases with the sample size, we get a nested sequence of parametric models, $\mathcal{H}_1^d \subset \mathcal{H}_2^d \subset \cdots$ which forms our sieve. Without restrictions, this sieve is very flexible and dense in the space of all continuous functions, as formalized in Theorem 2:

Theorem 2 (Universal Approximation of NG-CES DNN). *Consider the sieve*

$$\mathcal{H}_M^d = \left\{ H : [0,1]^d \to \mathbb{R} \,\middle|\, H(L;\theta) = \left(\sum_{m=1}^{M} \gamma_m \left[\left(\sum_{j=1}^{d} \gamma_{mj} \alpha_{mj}^{\rho_m} L_j^{\rho_m} \right)^{\tau_m/\rho_m} \right]^{\rho} \right)^{\tau/\rho}, \right.$$

$$\left. \text{with } \theta \in \Theta_M^d \equiv \left\{ \theta \in \mathbb{R}^{M(2d+3)+2} \,\middle|\, (\gamma_{mj}, \alpha_{mj}, \tau_m) \geq 0, \sum_{m=1}^{M} \alpha_{mj} = 1, \rho_m \leq 1 \right\} \right\}$$

Here, Θ_M^d removes redundancies and implements normalizations. Then, $\mathcal{H}^d = \bigcup_{M=1}^{\infty} \mathcal{H}_M^d$ is dense in the space of all continuous functions $C[0,1]^d$ under the sup-norm metric.

The intuition is that a polynomial of any order can be generated through a specific combination of sieve coefficients for a finite M. The result then follows from the polynomials being dense in the space of continuous functions as given by the Stone Weierstrass theorem.

C PROOFS

Proof of Theorem 1.

(i) First rewrite

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i(\alpha) = \frac{1}{n}\sum_{i=1}^{n} (\psi_i(\alpha) - \mathbb{E}[\psi_i(\alpha)]) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\psi_i(\alpha)]. \tag{37}$$

Using Assumption 1 empirical process theory from Andrews (1994),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (\psi_i(\alpha) - \mathbb{E}[\psi_i(\alpha)]) = O_p(1) \tag{38}$$

Using Assumption 2, uniformly in α

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \psi_i(\alpha)\right] \xrightarrow{p} m_1(\alpha). \tag{39}$$

Then by Assumption 39 and Assumption 3, uniformly in α and using $\lambda_n/n \to$

$\lambda_0 \geq 0$,

$$Z_n(\alpha) = \frac{1}{n} U_n(\alpha) = \left[\frac{1}{n}\sum_{i=1}^{n} \psi_i(\alpha)\right]' W_n(\alpha) \left[\frac{1}{n}\sum_{i=1}^{n} \psi_i(\alpha)\right] + \frac{\lambda_n}{n} \sum_{g=1}^{G} \left\|\alpha^{(g)}\right\|_2$$
$$\xrightarrow{p} m_1(\alpha)' W(\alpha) m_1(\alpha) + \lambda_0 \sum_{g=1}^{G} \left\|\alpha^{(g)}\right\|_2$$
$$= Z(\alpha).$$

So using (A.4) and Assumptions 1-3, uniformly in α,

$$\arg\min_{\alpha \in \Theta} Z_n(\alpha) \xrightarrow{p} \arg\min_{\alpha} Z(\alpha).$$

(ii) When $\lambda_n = o(n)$, (A.4) is modified in the following manner, uniformly in α :

$$Z_n(\alpha) = \frac{1}{n} U_n(\alpha) \xrightarrow{p} m_1(\alpha)' W(\alpha) m_1(\alpha).$$

So because α_0 is identified by Assumptions 2(a) and 3 and it is the unique minimum of the limit in (A.5), using Corollary 3.2.3 of Van der Vaart and Wellner (1996) we have the consistency result:

$$\widehat{\alpha}_n = \arg\min_{\alpha \in \Theta} Z_n(\alpha) \xrightarrow{p} \arg\min_{\alpha \in \Theta} \left[m_1(\alpha)' W(\alpha) m_1(\alpha)\right] = \alpha_0.$$

■

Proof of Theorem 2.

Consider the case of a scalar input $\ell \in [0, 1]$. Results for vector-valued input follow. We know that the coefficients are constrained to be in $\Lambda_d^{(K_1)}$ but let us restrict them further in the following way: For each node $j \in \{1, \ldots, K_1\}$ in the latent layer, let $A_j^{(1)} = 0$, $r_j^{(1)} = 1$ and $\tau_j^{(1)} = j$. Then, the output of each latent node j will be ℓ^j. Now, constrain the output layer to just be an affine transformation, so we have $r^{(2)} = 1$ and $\tau^{(2)} = 1$. The output of such a neural network with K_1 nodes in the hidden layer is of the form $h(\ell) = A^{(2)} + \gamma_1^{(2)} \ell + \cdots + \gamma_{K_1}^{(2)} \ell^{K_1} \equiv \sum_{j=0}^{K_1} \gamma_j^{(2)} \ell^j$ where $\gamma_0^{(2)} \equiv A^{(2)}$ and $\gamma_j^{(2)} \in \mathbb{R}$ for all j.

Now, consider any continuous function over the unit interval, $f \in C([0,1])$ and choose any target precision $\varepsilon > 0$. By the Stone–Weierstrass theorem, we know that there exist $K_1 < \infty$ and a vector of coefficients $\lambda \equiv \gamma^{(2)} \in \mathbb{R}^{K_1+1}$ (and a corresponding function $h(\ell,\lambda) \in \mathcal{H}_1$) such that the following holds:

$$\rho(f,h) \equiv \sup_{\ell \in [0,1]} \left| f(\ell) - h(\ell,\lambda) \right| < \varepsilon.$$

So, $\mathcal{H}_1$ is dense in the space of all continuous functions over $[0,1]$ when coefficients are unconstrained (or constrained to lie in $\widetilde{\Lambda}$). We can extend this to vector valued ℓ using the binomial expansion theorem. ■

D DETAILS OF THE DATASET

Table 8: Summary Statistics of Employment Hours of All Worker Types

	mean	std	min	25%	50%	75%	max
beds	99.87	48.46	1.00	66.00	96.00	120.00	1361.00
bed_days	35.19	17.67	0.05	22.26	33.58	43.80	496.76
assets	9.11	25.45	-20.66	1.11	2.25	5.43	242.86
grossrev	9.37	5.65	0.13	5.26	8.13	12.04	42.73
hrs_rn_donadmin	1.75	0.52	0.00	1.61	1.86	2.00	3.92
hrs_rnadmin	3.16	2.86	0.00	1.04	2.41	4.50	17.94
hrs_rn	9.37	6.80	0.00	4.52	7.77	12.49	47.05
hrs_lpn_admin	1.74	2.21	0.00	0.00	0.99	2.64	12.41
hrs_lpn	19.89	11.78	0.00	10.84	17.96	27.11	76.53
hrs_cna	52.54	26.62	0.00	32.61	47.94	67.47	189.84
hrs_na_trn	0.70	1.50	0.00	0.00	0.00	0.59	9.24
hrs_medaide	1.82	3.37	0.00	0.00	0.00	2.14	16.66
hrs_admin	2.46	2.37	0.00	1.67	1.93	2.10	17.29
hrs_meddir	0.09	0.12	0.00	0.00	0.05	0.13	0.87
hrs_othmd	0.02	0.08	0.00	0.00	0.00	0.00	1.43
hrs_pa	0.00	0.02	0.00	0.00	0.00	0.00	0.50

	mean	std	min	25%	50%	75%	max
hrs_np	0.03	0.12	0.00	0.00	0.00	0.00	1.31
hrs_clinnrsspec	0.00	0.03	0.00	0.00	0.00	0.00	0.67
hrs_pharmacist	0.08	0.08	0.00	0.01	0.07	0.12	1.04
hrs_dietician	0.60	0.75	0.00	0.10	0.30	0.83	6.40
hrs_feedasst	1.23	3.77	0.00	0.00	0.00	0.00	22.56
hrs_ot	1.70	1.51	0.00	0.49	1.42	2.27	9.73
hrs_otasst	2.16	1.76	0.00	0.85	1.82	3.12	10.77
hrs_otaide	0.02	0.17	0.00	0.00	0.00	0.00	2.28
hrs_pt	1.71	1.55	0.00	0.50	1.40	2.29	10.39
hrs_ptasst	2.54	1.96	0.00	1.14	2.08	3.62	12.38
hrs_ptaide	0.25	0.58	0.00	0.00	0.00	0.02	3.86
hrs_respther	0.10	0.64	0.00	0.00	0.00	0.00	10.28
hrs_resptech	0.00	0.05	0.00	0.00	0.00	0.00	2.17
hrs_spclangpath	1.21	0.98	0.00	0.39	1.04	1.77	5.42
hrs_therrecspec	0.10	0.41	0.00	0.00	0.00	0.00	3.58
hrs_qualactvprof	1.37	0.99	0.00	0.36	1.68	1.93	6.65
hrs_othactv	2.65	2.70	0.00	0.42	1.96	3.89	16.34
hrs_qualsocwrk	1.37	1.20	0.00	0.02	1.57	1.95	7.26
hrs_othsocwrk	0.74	1.08	0.00	0.00	0.00	1.54	5.73
hrs_mhsvc	0.01	0.09	0.00	0.00	0.00	0.00	2.36

E COMPUTATION

To estimate our model of team-based production, we encode this as a custom neural network architecture that leverages momentum-based projected stochastic gradient descent. Neural networks are chosen for their computational efficiency and ability to handle complex, high-dimensional data. This allows us to flexibly model intricate substitution patterns among different worker types while maintaining computational feasibility.

We begin by initializing the neural network parameters. The network architec-

Table 9: Correlation Matrix of Various Variables

	adm	**nurse**	**spcl**	**well**	**ther**	**assets**	**grev**	**beds**	**bed days**
adm	1.00	0.51	0.15	0.38	0.47	0.04	0.44	0.12	0.37
nurse	0.51	1.00	0.21	0.50	0.64	0.09	0.76	0.22	0.64
spcl	0.15	0.21	1.00	0.22	0.23	0.03	0.23	0.04	0.13
well	0.38	0.50	0.22	1.00	0.31	0.02	0.42	0.12	0.35
ther	0.47	0.64	0.23	0.31	1.00	0.11	0.68	0.13	0.40
assets	0.04	0.09	0.03	0.02	0.11	1.00	0.21	0.04	0.10
grev	0.44	0.76	0.23	0.42	0.68	0.21	1.00	0.20	0.60
beds	0.12	0.22	0.04	0.12	0.13	0.04	0.20	1.00	0.32
bed days	0.37	0.64	0.13	0.35	0.40	0.10	0.60	0.32	1.00

After data cleanup (e.g., removal of outliers, missing values), we have data for 8,144 nursing homes spanning three years (2017-2019), yielding a balanced panel with 24,432 observations.

ture consists of two layers with a CES (constant elasticity of substitution) activation function. The initial parameters for each node, such as γ, τ, r, and A, are set according to predefined initial values. We utilize the Adam optimizer for efficient training, initializing moving averages for gradients ($m_0 = 0$) and squared gradients ($v_0 = 0$), and setting the timestep to $k = 0$.

For each simulation, we generate a dataset $\mathcal{D}$ using a data-generating process (DGP) characterized by parameters n, β, and σ_ϵ. The dataset includes observations $\{\ell_i, y_i\}_{i=1}^n$. During each epoch of training, we sample a mini-batch $\mathcal{B}$ from the dataset, introducing stochasticity, which helps the model generalize better by training on random subsets of data.

In the forward pass, the neural network computes predicted outputs y_{pred} for all sampled observations by processing the input data through its layers. This involves applying the CES activation function and combining the input features according to the network's current parameters. The predicted outputs are then used to compute the penalized and shape-constrained sample Sieve-GMM objective function:

$$\mathcal{L} \equiv \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\alpha)\right]' W_n(\alpha)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\alpha)\right] + \lambda_n \sum_{g\in\mathcal{G}^*}\sqrt{p_g}\left\|a_{g\cdot}\right\|_2 .$$

In the backward pass, we calculate the subgradients of the objective function $\mathcal{L}$ with respect to the model parameters Θ. These subgradients indicate how much each parameter needs to be adjusted to minimize the objective function. The Adam optimizer updates the parameters by first adjusting the moving averages of the gradients and squared gradients using the decay rates β_1 and β_2:

$$m_k \leftarrow \beta_1 m_{k-1} + (1-\beta_1)\partial_\Theta \mathcal{L}, \quad v_k \leftarrow \beta_2 v_{k-1} + (1-\beta_2)(\partial_\Theta \mathcal{L})^2$$

These moving averages are then corrected for bias:

$$\hat{m}_k \leftarrow \frac{m_k}{1-\beta_1^k}, \quad \hat{v}_k \leftarrow \frac{v_k}{1-\beta_2^k}$$

The parameters are updated using the corrected moving averages, scaled by the learning rate α:

$$\Theta \leftarrow \Theta - \alpha \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \epsilon}$$

After each update, the parameters are projected onto a feasible set $\mathcal{C}$ to enforce the parameter constraints and normalizations discussed before:

$$\Theta \leftarrow \mathrm{Proj}_{\mathcal{C}}(\Theta)$$

After training epochs are completed, we draw a new sample $\mathcal{D}'$ from the DGP for evaluation purposes. The model's performance is assessed by computing the root mean square error (RMSE) on this new sample:

$$RMSE_h = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(y_i' - h(\ell_i'; \Theta))^2}$$

We repeat this process across multiple simulations, storing the RMSE and coefficients for each run. Finally, we compute the average RMSE and average coefficients across all simulations:

$$RMSE_{\Theta_j} = \sqrt{\frac{1}{S}\sum_{sim=1}^{S}(\Theta_{j,sim} - \overline{\Theta}_j)^2}$$

F SUPPLEMENTAL TABLES AND FIGURES

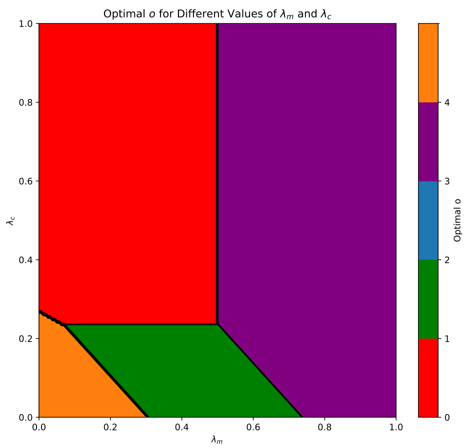


Figure 6: Linear Cost

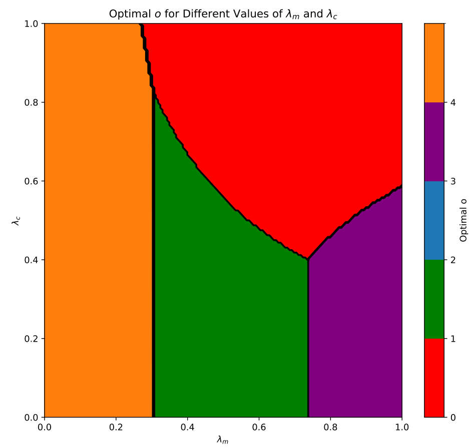


Figure 7: Lasso-type cost

Figure 8: Firm's Optimal Organisational Structure under Different Costs

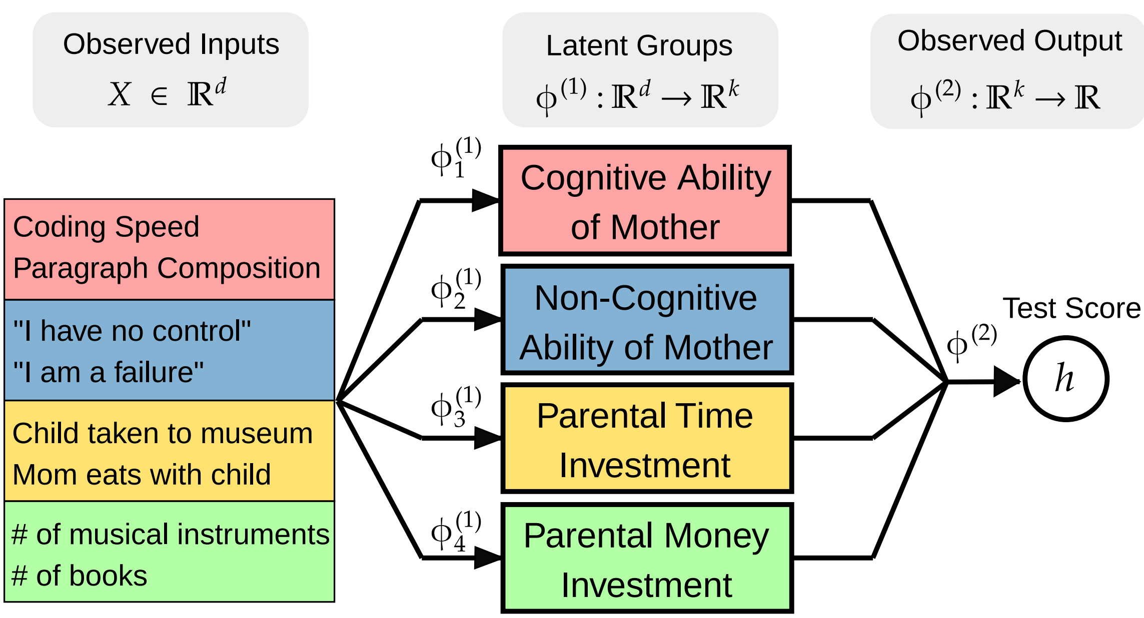
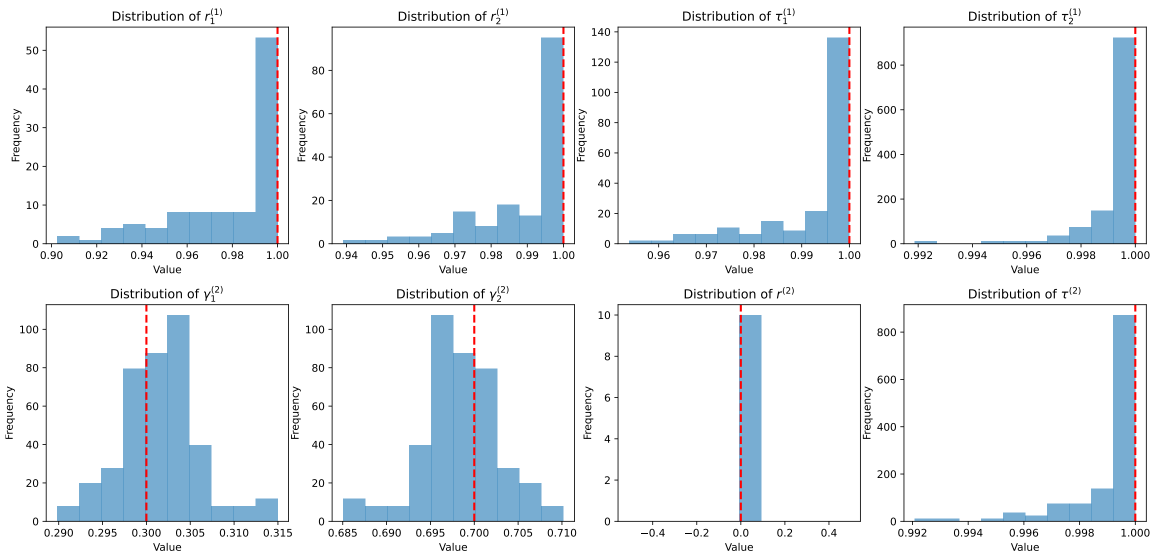


Figure 9: Skill Formation Model Grouping Structure

Table 10: Point Estimates and RMSE of Parameters in Nested CES Monte Carlo

Type	Coefficient	True Value	Average Estimate	RMSE
Team 1	$\gamma_{11}^{(1)}$	1	0.99	3×10^{-5}
	$\gamma_{12}^{(1)}$	1	0.99	6×10^{-5}
	$\gamma_{13}^{(1)}$	0	3×10^{-6}	3×10^{-5}
	$\gamma_{14}^{(1)}$	0	2×10^{-5}	2×10^{-4}
	$r_1^{(1)}$	1	0.98	0.03
	$\tau_1^{(1)}$	1	0.99	0.01
Team 2	$\gamma_{11}^{(2)}$	0	3×10^{-6}	4×10^{-5}
	$\gamma_{12}^{(2)}$	0	8×10^{-6}	6×10^{-5}
	$\gamma_{13}^{(2)}$	1	0.99	3×10^{-5}
	$\gamma_{14}^{(2)}$	1	0.99	2×10^{-4}
	$r_1^{(2)}$	1	0.99	0.01
	$\tau_1^{(2)}$	1	0.99	0.001
Team 3	$A^{(2)}$	1	0.99	0.008
	$\gamma_1^{(1)}$	0.3	0.30	4×10^{-3}
	$\gamma_2^{(1)}$	0.7	0.69	4×10^{-3}
	$r^{(2)}$	0	-0.0	4×10^{-10}
	$\tau^{(2)}$	1	0.99	0.001

Figure 10: Non-Standard Sampling Distribution with Binding Shape Constraints



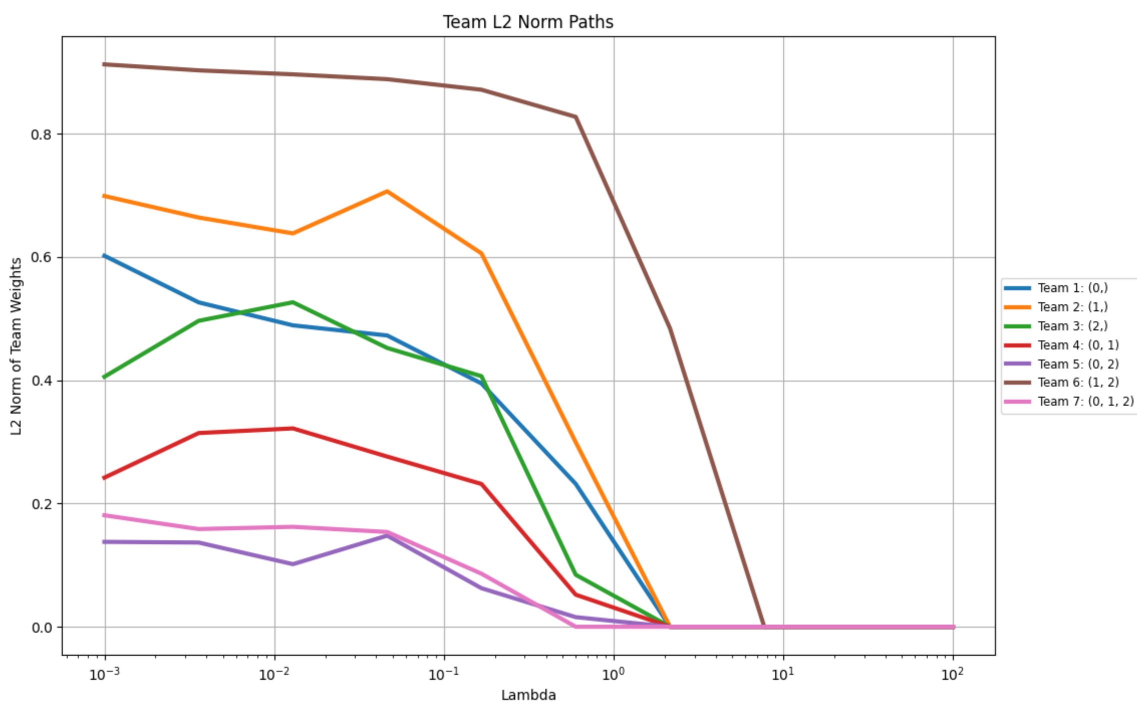


Figure 11: Configuration Paths Under Soft Time Budget (Inequality) Constraint

Figure 12: Worker-Specific Parameter Plots (OLS)

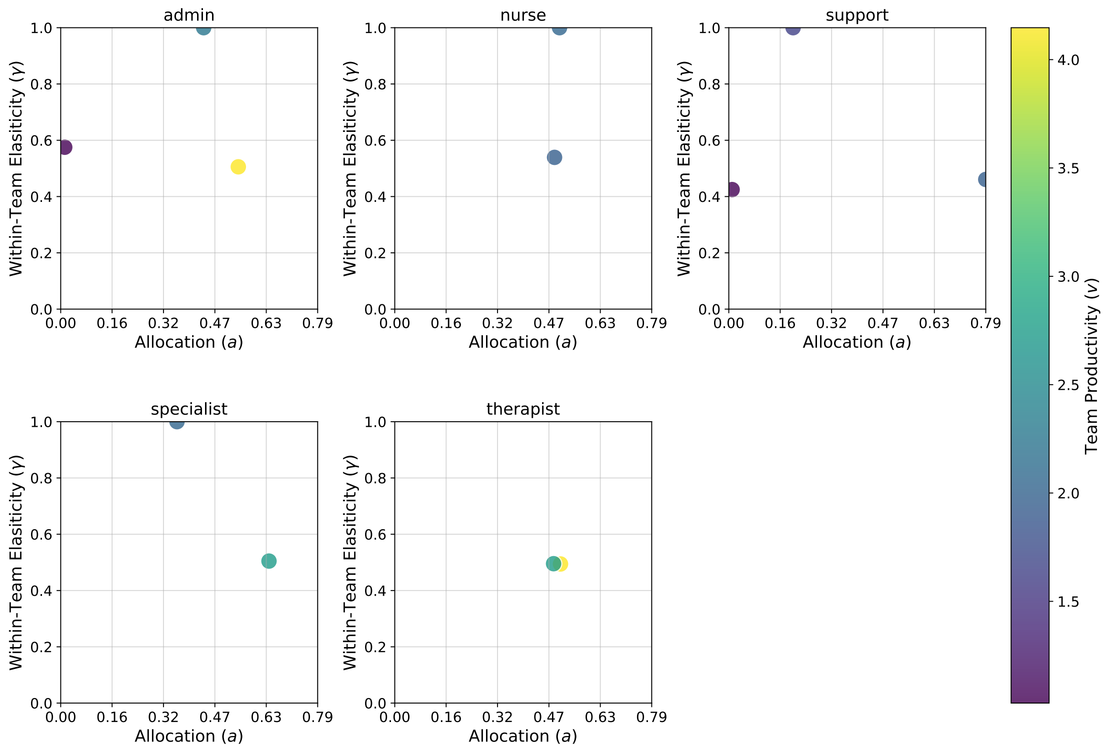


Figure 13: Team Hypergraph (GMM)

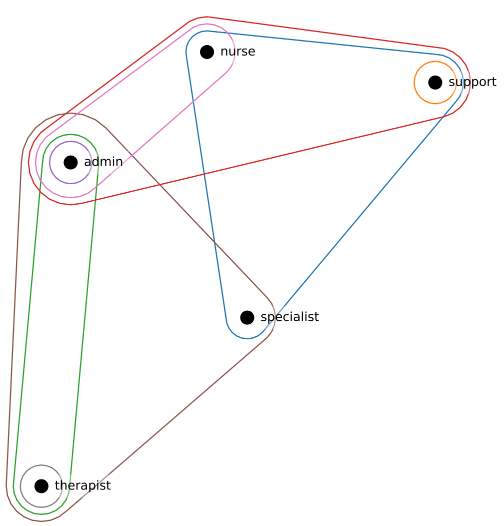


Figure 14: Team Hypergraph (OLS)

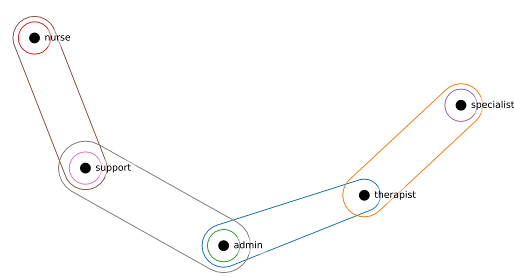


Table 11: List of Teams Eliminated Under OLS and GMM

S. No.	Eliminated Under	Team
1	GMM	Nurse
2	GMM	Specialist
3	OLS	Therapist
4	OLS	Admin, Nurse
5	Both	Admin, Specialist
6	GMM	Admin, Wellness
7	GMM	Nurse, Wellness
8	Both	Nurse, Specialist
9	Both	Nurse, Therapist
10	Both	Wellness, Specialist
11	Both	Wellness, Therapist
12	GMM	Specialist, Therapist
13	OLS	Admin, Nurse, Wellness
14	Both	Admin, Nurse, Specialist
15	Both	Admin, Nurse, Therapist
16	Both	Admin, Wellness, Specialist
17	Both	Admin, Wellness, Therapist
18	OLS	Admin, Specialist, Therapist
19	OLS	Nurse, Wellness, Specialist
20	Both	Nurse, Wellness, Therapist
21	Both	Nurse, Specialist, Therapist
22	Both	Wellness, Specialist, Therapist
23	Both	Admin, Nurse, Wellness, Specialist
24	Both	Admin, Nurse, Wellness, Therapist
25	Both	Admin, Nurse, Specialist, Therapist
26	Both	Admin, Wellness, Specialist, Therapist
27	Both	Nurse, Wellness, Specialist, Therapist
28	Both	Admin, Nurse, Wellness, Specialist, Therapist